The Dissertation Committee for Yang-Chi Chen
certifies that this is the approved version of the following dissertation:

Knowledge-based Learning for Classification of Hyperspectral Data

Committee:

Melba M. Crawford, Supervisor

Joydeep Ghosh, Supervisor

Paul Damien

David Morton

Elmira Popova

Knowledge-based Learning for Classification of Hyperspectral Data

by

Yang-Chi Chen, B.S., M.S.E.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2007

To My Family

Knowledge-based Learning for Classification of Hyperspectral Data

Publication No. _____

Yang-Chi Chen, Ph.D.
The University of Texas at Austin, 2007

Supervisor: Melba M. Crawford
Supervisor: Joydeep Ghosh

This research focuses on three critical issues related to land cover classification using hyperspectral data: i) robust classification of high dimensional input data; ii) utilization of contextual spatial information; and iii) knowledge transfer for classification of data for which little or no labeled samples are available.

An integrated max-cut hierarchical decomposition algorithm that uses support vector machines to classify multi-class land cover data is proposed to address the high dimensional input problem. The hierarchical support vector machine (HSVM) classifier solves a series of max-cut binary set partitioning problems to hierarchically and recursively partition the set of classes into two subsets until pure leaf nodes are

obtained. Support vector machines are used at each internal node of the hierarchy to construct the binary decision boundary. It is shown to perform well with limited amount of ground truth.

Although hyperspectral data provide new capabilities for discriminating spectrally similar classes, it is often useful to incorporate reliable spatial information. A knowledge-based stacking approach is proposed to utilize spatial information within homogeneous regions and at class boundaries. The proposed max-cut HSVM approach (MC-HSVM) learns the location of the class boundary and combines original bands with the extracted spectral information of a neighborhood to train the HSVM classifier. An ensemble of majority filtering and MC-HSVM is also investigated to handle complex spatial neighborhoods through a switch process.

Since the spectral signatures could be affected by many uncontrollable factors, a classifier must capture the resulting variations in spectral signatures. Inspired by nonlinear manifold learning, a shortest path k-nearest neighbor classifier (SkNN) is proposed for the analysis of spatially disjoint data and multi-temporal images. The ability to update an existing model so that it performs well on images with no labeled data leads to many potential applications of land cover classification. As a result, this research simplifies the land cover classification process and increases the accessibility of hyperspectral sensors through the development of intelligent classification algorithms.

Algorithms proposed in this research help solve the three critical problems outlined previously and achieve the objective of this study: to develop efficient, knowledge-based classification procedures for hyperspectral sensed image data.

Acknowledgments

I would like to thank my supervisor, Dr. Melba Crawford, for allowing me to learn and explore new ideas and being a mother figure who push me to get better in every aspects. I also want to thank my co-supervisor, Dr. Joydeep Ghosh, for taking me under his wings after Dr. Crawford was promoted as the Director of LARS at Purdue University. I treasure my time spent with my friends here at Austin. They help me realize that there is life beyond studying. I am really thankful for having a family that gives me their constant support. My parents allowed me to live in another country that is thousands of miles away from home and could only see me for less than once a year. My wife gave up her job in Taiwan so she could be here with me and became my "PhD assistant". I will not be able to finish this dissertation without them.

Yang-Chi Chen

The University of Texas at Austin

May 2007

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Over the last thirty years, researchers have been using remotely sensed data to study the earth's ever-changing land cover. In this chapter, supervised classification processes and their applications in land cover mapping using the recently available spaceborne hyperspectral data are introduced. To improve performance and increase the utilization of this important research, this study proposes a series of new classification processes that to reduce computation times and lower the knowledge threshold required to utilize hyperspectral data, thereby reducing the requirement of human expertise, which maintaining good classification accuracies.

1.1 The Classification Problem

Classification algorithms map a potentially large input space (attributes) to a single-dimensional label via a collection of hypotheses. Widely used in both business and research applications, classifiers help systematize the decision making process. For example, character recognition uses classification algorithms to identify written or printed characters in a document. As applied by the U.S. Postal Service, this process speeds up the mail sorting process, shortens mail delivery time, and reduces both the

physical work load and postal worker exposure to hazards. Similarly, classification is employed in a wide variety of applications including bio-informatics, web searches, and remote sensing image analysis.

1.1.1 Background

To solve classification problems such as automatic character recognition problems, the classification algorithm must be capable of facilitating good decision making and of performing efficiently. Classification algorithms are categorized as supervised or unsupervised, according to the information that is available to train the method to a specific problem. Supervised methods have labeled data available to train the algorithm, while unsupervised methods do not. As such, unsupervised methods seek to identify homogeneous groups/clusters of data, rather than assign a particular class label. Supervised methods seek to label objects in accordance with the characteristics of the input data and can be evaluated relative to their capability to correctly classify labeled data. As such, supervised methods are preferred where possible.

Two stages are commonly involved in solving a supervised classification problem: "training" and "testing." In the training stage, the classifier learns and modifies its classification models according to previous experiences (training samples). In the testing stage, a novel observation is provided to the classification model and is labeled according to the algorithm's final result. For example, the postal zip code recognition process involves "training" a classifier using data representing the hundreds of ways people write numbers. The classifier then develops a model through measurable attributes such as the maximum width, height, average width, and correlation of height and width of each digit. In the testing stage, the trained classifier measures and classifies new samples, and the results are used to evaluate a classifier's performance.

Supervised classification algorithms help solve real world problems such as

character recognition, cancer detection, and land cover classification. Most land cover classification algorithms are categorized as supervised classifiers that are trained by ground truth (labeled samples) collected from accessible regions. The learned model is applied to classify the entire extended region.

1.1.2 Land Cover Classification

> Mapping Earth's Diverse Landscape: "Nearly every aspect of our lives is tied into the vegetation and ground cover that surround us. Farms feed us, forests provide us with oxygen and building materials, rivers and lakes yield fresh water to drink, and cities shelter us. When land covers change, our health, economy, and environment can all be affected."
>
> "For years, scientists across the world have been mapping changes in the landscape to prevent future disasters, monitor natural resources, and collect information on the environment." *–NASA Earth Observatory https://earthobservatory.nasa.gov/Library/LandCover/*

Land cover mapping is one way of measuring environmental changes, and thus provides valuable information. The goal of land cover mapping is to identify "classes" of cover on the earth's surface, e.g., water, soil or specific vegetation types over an extensive geographical region. Because of cost and geographical access, it is almost impossible to classify land cover over large remote areas by conducting ground surveys. However, aircraft and satellites equipped with sensors can now acquire large amounts of data that measure land cover characteristics efficiently and accurately, thereby making data acquisition for land cover classification problems a practical reality for extended regions.

Passive optical sensors used for natural land cover classification are categorized according to their number of spectral bands: multispectral and hyperspectral. Multispectral sensors, which sense integrated responses over specified intervals (10s

- 100s of nm) of the electromagnetic spectrum as discrete "band" values, have been in operation for over 30 years. Since each multispectral sensor is designed to cover specific broad wavelength ranges, it has a limited set of applications. In the last decade, there has been an increase in the availability of data from hyperspectral sensors which simultaneously acquire hundreds of bands defined over narrow (5-10 nm), contiguous wavelength ranges. These sensors provide detailed chemical information that closely approximates the continuous response from a target across a range of wavelengths, which, in turn, contributes to improved classification accuracy. Hyperspectral sensors provide a rich set of spectral information for recognizing natural land cover types such as water, soil, minerals and vegetation at the species level. Many researchers use this type of sensor to analyze eco-systems, urban landscapes, and the impact of natural disasters.

1.1.3 Land Cover Classification using Hyperspectral Sensors

Airborne sensors, such as the Hyperspectral Digital Imagery Collection Experiments (HYDICE) and the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) have been used extensively for land cover classification studies. Figure 1.1 depicts AVIRIS[1] aboard an airplane, scanning the earth's surface and acquiring 224 bands simultaneously. The bands provide continuous coverage of wavelengths from 400nm to 2400nm. Experiments that use data from these airborne sensors show great improvements in results relative to using traditional multispectral data in classification [59]. However, the availability and high cost of flying airplanes that carry these instruments over specific research areas limits the accessibility to data acquired by these sensors and makes it almost impossible to conduct multi-temporal studies.

With the launch of the Hyperion sensor on the NASA Earth Observation-1 (EO-1) satellite (See Figure 1.2 and *http://eo1.gsfc.nasa.gov/miscPages/home.html*),

[1] http://aviris.jpl.nasa.gov/

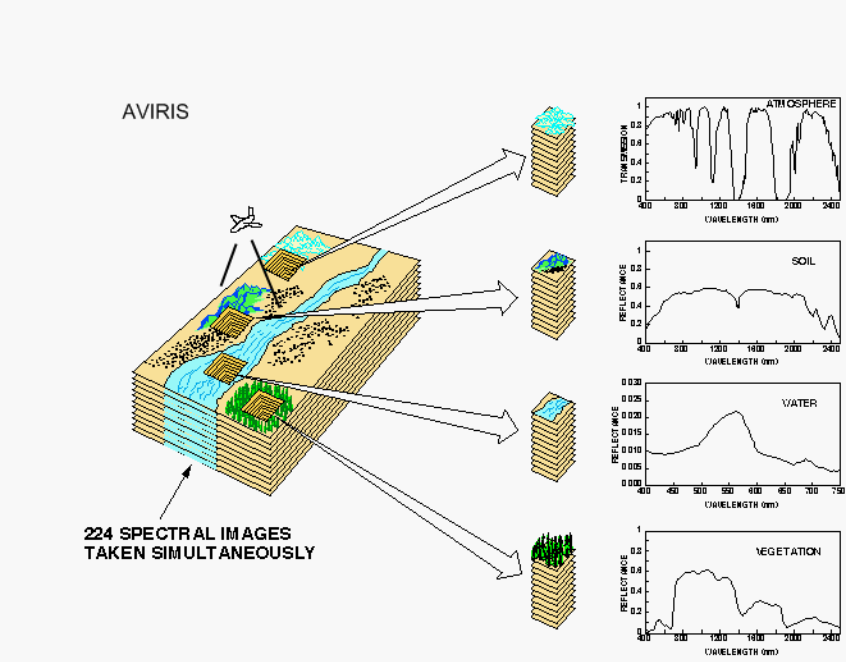

Figure 1.1: Hyperspectral Data from the NASA AVIRIS airborne sensor

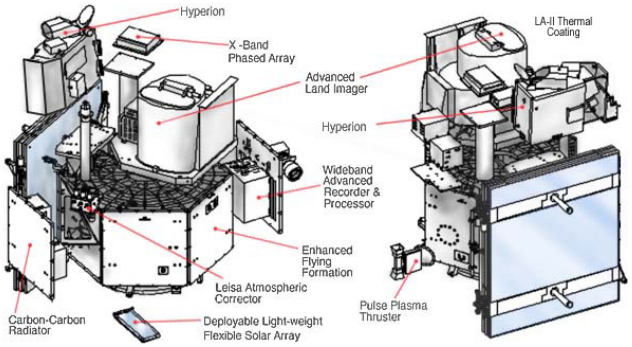

Figure 1.2: Instruments on the Earth Observation-1 Satellite

researchers in the remote sensing community could begin to exploit data collected by hyperspectral sensors over extended areas and in multiple time periods at minimal cost. Hyperion has two sets of detectors that record data in 220 bands over the visible and near infrared (VNIR) and short wavelength infrared (SWIR) regions of the spectrum. Each Hyperion image covers a 100km*7.5km strip of data with a 30m*30m resolution. This sensor has served the remote sensing community since 2001 and has provided a tremendous amount of data at a relatively low cost for land cover studies.

Although hyperspectral sensors provide much more information than traditional multispectral sensors, they also present challenges. The high dimension of the input data necessitates estimation of a very large number of parameters for commonly used statistical classification methods. Because of the inaccessibility of many regions, the number of labeled samples (ground truth) available to researchers is small relative to the high dimensionality of the features (input space.) Also, a large number of classes (output spaces), which is typical to land cover classification, often require complex decision boundaries in the classification model.

1.2 Motivation

Hyperspectral sensing is a relatively new technology that has tremendous potential, but is under utilized, partially because of the lack of readily available methods which can be used reliably by the applications community. Overall, this research is motivated by the desire to simplify the land cover classification process and increase the accessibility of hyperspectral sensors through the development of intelligent classification algorithms which require limited expertise by users. This research addresses the problem in three ways. Firstly, it seeks to provide users with a classifier that handles high dimensional inputs, while supplying robust classification results. This classifier must be efficient, require a minimum amount of tuning, and yield mean-

ingful results that are easy to understand. Secondly, the research seeks to enhance basic spectral-based classification algorithms and to utilize contextual information, which is robust but difficult to extract, in the classification process. Thirdly, it seeks to solve the difficult problem of classifying multiple images acquired from areas that have limited or possibly no labeled samples. Because spectral signatures of land cover vary over extended areas due to many uncontrollable factors, it is critical to develop a classifier that can adapt to such spectral changes.

1.2.1 Limited Number of Labeled Samples

Hyperspectral data sets are extremely large, and many input features are highly correlated and possibly redundant. High dimensional inputs are problematic for statistical classification problems that involve estimation of the covariance matrix, because sample sizes are often small due to the inaccessibility of many areas. This problem is referred as "the curse of dimensionality." Many feature selection and feature projection approaches have been proposed to remove or aggregate redundant features to maximize discrimination between classes [41]. These methods usually provide accurate classification on training data when sample sizes are small, but they often fail to do so when they classify pixels from areas that have slightly different spectral signatures [18]. These methods often overtrain the models and lose the information provided by the original bands when the number of features is reduced. A loss in diversity results in poor transfer of knowledge. Previous studies performed by the UT Remote Sensing Group [2] demonstrated that ensemble methods and nonparametric classifiers maintain their diversity while providing good classification accuracies on the original training and testing data set. The weakness of these methods is that they require a long, careful tuning process. A new computationally efficient classification algorithm is proposed in this study to handle high

[2]The Remote Sensing Group at the Center for Space Research of the University of Texas at Austin

dimensional input and small training sample size problems, while preserving the diversity of the classifier.

1.2.2 Utilize Spatial Information

The Hyperion sensor on the EO-1 satellite has the advantage of representing spectral signatures in much greater detail than traditional multispectral spaceborne sensors (which only have ∼3-15 bands), and thus has much greater potential for providing improved characterization and discrimination of targets. Although Hyperion is able to collect hundreds of bands simultaneously, calibration is difficult because it is a pushbroom sensor, and the signal-to-noise ratio is low for certain wavelengths resulting in "striped" columns in many bands. While normalization of statistics in local windows and the application of low pass filters can mitigate the effect, these approaches are often inadequate and can even induce artificial effects in the data. Spatial neighborhood information, which is often more reliable but difficult to analyze, provides an alternative source of information which should be utilized in conjunction with spectral data to label the class pixels.

Most land cover classification approaches focus on pixel-wise classification that only utilizes spectral information associated with a given pixel location. These approaches ignore some important characteristics of geographical data, including region shape, location, and relation of neighboring samples to the targeted sample. This information is recognized as spatial information. It is often not well represented in the labeled images and requires additional understanding of the survey site to identify the correct land cover type. For instance, it is impossible to explore classes in neighborhoods where they cannot physically exist adjacent to each other via a pixel-wise classification. It is also difficult to flag outliers in a homogeneous area using the traditional pixel labeling schemes.

Classification procedures that can learn and utilize both spatial and spectral

information could potentially dramatically reduce the time required to process the image and better utilize the time and knowledge of an expert. Previous studies using multispectral sensors such as Markov random fields (MRF) or simple stacked vector approaches for multispectral data are problematic for hyperspectral data because of the curse of dimensionality and the increased complexity of image texture represented by the detailed spectral signatures.

1.2.3 Knowledge Transfer Problems

Traditionally, the training and test data are spatially co-located and can thus be assumed to be samples from the same distribution. Because of the difficulties in physically accessing certain areas or limitation of time and budget, it is difficult to have enough ground truth data from a new region for land cover classification. Thus, it is also useful to evaluate classifier performance when applied to areas that are somewhat removed from the original training and test area. These new test samples could be from a spatially disjoint area of the same image or from images acquired by the same sensor, but at a different time or date. Because the spatial signatures are affected by sun angle, atmospheric conditions and acquisition times, the spectral reflectance of land cover can vary, often nonlinearly, from image to image.

Consider an example from the Okavango Delta of Botswana. Plots of average spectral responses of both water and hippo grass are shown in Figure 1.3. This figure demonstrates why hippo grass of image 2 can be easily misclassified as water if the model learned from image 1 is used to classify image 2. Since hippo grass grows along the river channel, its density, as well as the sun and acquisition angles of the satellite can cause changes in spectral signatures shown in Figure 1.3 and result in classification errors. To accommodate changes of distributions of certain land cover types, a new process must be developed to preserve knowledge learned from the

original area while evolving with changes in new labeled samples.

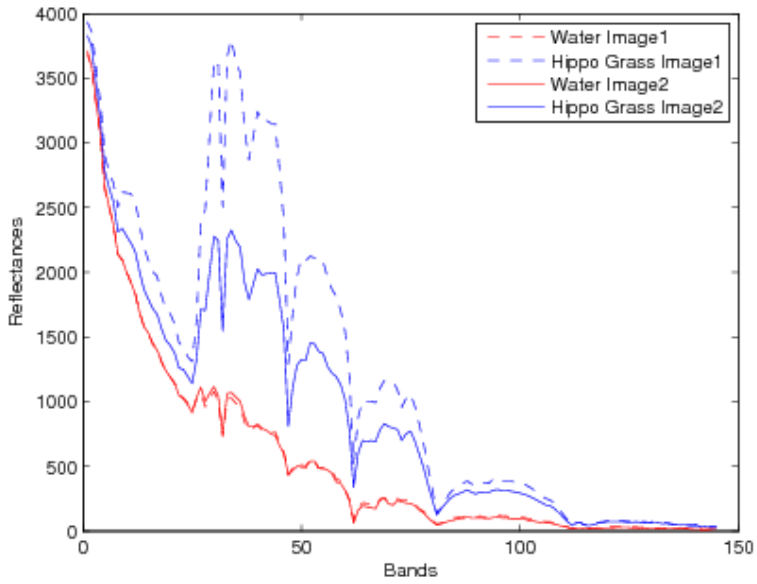

Figure 1.3: Multitemporal Spectral Signatures

This land cover classification study focuses on the utilization of space-based hyperspectral data. Overcoming potential obstacles such as an insufficient number of samples for estimating multivariate statistics and the shortcomings of traditional contextual classification algorithms in handling the hyperspectral data provides the main motivation of this study.

1.3 Problem Statement: Efficient, Knowledge-based Classification Procedures for Hyperspectral Remotely Sensed Data

The importance and difficulties of land cover classification using hyperspectral data are discussed in the first part of this chapter. This study involves development of a

series of advanced approaches that will hopefully inspire more users to avail themselves of the full potential of the rich spectral information inherent in hyperspectral image data. To achieve this difficult task, three interrelated problems are addressed in Chapters 3, 4 and 5 of this study.

1. Statistical classification of hyperspectral data is challenging because the inputs are high in dimension and represent multiple classes that are sometimes quite mixed, while the amount and quality of ground truth in the form of labeled data are typically limited. The resulting classifiers are often unstable and have poor generalization. An integrated max-cut hierarchical decomposition algorithm that uses support vector machines to classify multi-class land cover data is proposed to address this problem. The new algorithm, referred to as the hierarchical support vector machine (HSVM), solves a series of max-cut binary set partitioning problems to hierarchically and recursively partition the set of classes into two subsets until pure leaf nodes are obtained. Support vector machines are used at each internal node of the hierarchy to construct the binary decision boundary.

2. Hyperspectral data provide new capabilities for discriminating spectrally similar classes, but unfortunately such class signatures often overlap in multiple narrow bands. Thus, it is useful to incorporate reliable spatial information when possible. A new knowledge-based stacking approach is proposed to utilize spatial information within homogeneous regions and at class boundaries, while avoiding the curse of dimensionality. The proposed max-cut HSVM approach (MC-HSVM) learns the location of the class boundary and combines the original bands with the extracted spectral information of a neighborhood to train a hierarchical support vector machine (HSVM) classifier. In addition, an ensemble of a majority filter-based approach and MC-HSVM is also presented in this study to handle complex spatial neighborhoods through a

weighted voting process.

3. Because the spectral signatures could be affected by many uncontrollable factors such as sun angle, atmospheric conditions and acquisition times, a classifier that updates its model dynamically is required to capture the differences. Inspired by nonlinear manifold learning, the shortest path k-nearest neighbor classifier (SkNN), which utilizes nonlinear manifold learning, is proposed for the analysis of spatially disjoint data and multi-temporal images. The ability to update an existing model so that it performs well on images for which there is no labeled data could contribute to many potential applications of land cover classification, particularly for hyperspectral data.

The background of the topics discussed in this chapter are reviewed in Chapter 2: Background and Related Work. A new hierarchical decomposition method is presented in Chapter 3. Contextual information and manifold learning that help solve knowledge transfer problems are explored in Chapter 4 and Chapter 5, respectively. Conclusions of the study are contained in Chapter 6, which also provides suggestions for further investigation. Additional information regarding the detailed discussion of the study sites over which data were acquired is contained in Appendix A. The ensemble study, which was conducted during this study and is repeatedly referred to in this research, is included in Appendix B of this dissertation.

Land cover classification using hyperspectral sensors is an important application of remote sensing. Three critical difficult problems which impede widespread application of these data have been presented. Advances in these three problem areas should help achieve the objective of this study: to develop efficient, knowledge-based classification procedures for hyperspectral remotely sensed image data.

Chapter 2

Background and Related Work

The goal of this chapter is to provide sufficient background to support the proposed research in land cover classification using hyperspectral sensors. As stated in the previous chapter, the proposed research in knowledge-based classification must deal with three problems: large input and output spaces with a limited amount of labeled data, contextual classification and the knowledge transfer problem. Previously developed approaches used to handle these issues are reviewed in this chapter.

2.1 Small Sample Sizes and High Dimensional Input and Output Space Problems

The majority of this chapter focuses on the small sample size problem, one of the most critical issues that researchers face when analyzing hyperspectral data for land cover classification. Most studies emphasize reducing the number of inputs through feature selection and feature extraction or the implementation of nonparametric classifiers e.g., support vector machines (SVM). Details of these approaches are presented in Section 2.1.1. In addition, the UT Remote Sensing Group, of which I am a member, proposed a binary hierarchical classifier (BHC) that uses Fisher's linear

discriminant and the generalized associative modular learning system (GAMLS) to handle large output space problems. These algorithms are discussed in Section 2.1.2. The UT remote sensing group also explored the idea in [34] that ensemble methods can mitigate the limited labeled sample problem and provide good classification of training and testing samples. The idea is briefly introduced in this chapter, while details of this study are shown in Appendix B.

2.1.1 High Dimensional Input Space Problems

Various approaches have been investigated to mitigate the impact of small sample sizes and high dimensionality, which are inherently coupled issues since the adequacy of a data sample depends on the data dimensionality, among other factors [66]. For example, regularization methods try to stabilize the covariance matrix by weighting the sample covariance matrix and a pooled covariance matrix or by shrinking the sample covariance matrix toward the identity matrix [73]. While this may reduce the variance of the parameter estimates, the bias of the estimates can increase dramatically. Several studies augment the small training set with unlabeled data and use semi-supervised learning techniques. These methods have been shown to enhance supervised classification [39, 71]. However, convergence of the updating scheme can be problematic and is affected both by the selection of the initial training samples and outliers. Alternatively, the input space can be transformed into a reduced feature space via feature selection [71] or feature extraction. Although these two approaches reduce the effect of the high dimensionality problem, feature selection methods are often trapped in a local optimal feature subset, while feature extraction methods lose the interpretability of the original features. Other approaches for dealing with a smaller labeled set are based on nonparametric classifiers such as decision trees (DT) [61, 10] or SVM [69]. These methods do not estimate the probability distribution of samples and tend to be able to avoid the

curse of dimensionality. More details of feature space representation and SVM are presented in sections that follow.

Feature Space Representation

The first challenge to any classification methodology is its input (feature) space representation. Although input features provide valuable information to classifiers, they can also include redundant or misleading information. In this section, methods that extract useful information from the original feature space are reviewed and categorized according to whether they are feature selection or feature extraction approaches.

Feature Selection algorithms select a subset of d' from the original d dimensional input, with the ultimate goal of achieving improved parameter estimates and good classification accuracy. This approach poses two challenges. First, identifying the best d' is an NP-hard combinatorial problem that may have a large search space. Second, finding the most appropriate objective function for this combinatorial problem is difficult because the objective function does not guarantee success in achieving good classification accuracy. To deal with the computational challenge, most approaches rely on a heuristic search through the feature space by adding and deleting individual features from a subset. Heuristic searches include greedy feature selection such as forward selection, backward elimination and non-greedy selection methods such as tabu-search [44].

Two approaches are used to evaluate the best feature subset. The first treats the classifier as a "black box" and uses an external loop that systematically adds and deletes features from the feature subset. The feature subset is evaluated according to its classification accuracy or relative to some distance measure relating the multivariate distributions of the classes. Because this approach builds a new classifier each time a feature is added or deleted from the feature subset, the training

process is very slow. The second approach focuses on creating objective function indices such as the Gini measure or entropy [61]. These methods are computationally superior, but suffer from the problem that no one general measure works for all classifiers.

Although subset selection methods may provide valuable domain knowledge about the importance of inputs, they are sensitive to anomalies in the training data. When the training sample size is small, they may not yield robust classifiers with good generalization. A recently developed feature selection approach involves selecting a random subspace of the original features as inputs to each classifier in an ensemble, thereby constructing multiple classifiers in the resulting random input space [37]. This approach is discussed in conjunction with other ensemble methods in Section 2.1.3.

Feature Extraction methods transform the original feature space to a much smaller set of new features via some linear or nonlinear functions. While this may result in some loss of interpretability, it reduces the original combinatorial problem into an easier optimization problem. Extraction methods such as principal component analysis and maximum noise fraction, which are referenced in Chapter 5, are discussed in this context in the remainder of this section.

- Principal Component Analysis: Principal component analysis (PCA) represents the covariance structure of a set of variables through a reduced number of orthogonal linear combinations of the variables [1, 43]. The original feature set **X'** $= [X_1, X_2, ..., X_d]$. Consider the new linear combinations

$$\begin{aligned} Y_1 &= a_{11}X_1 + a_{12}X_2 + ... + a_{1d}X_d \\ Y_2 &= a_{21}X_1 + a_{22}X_2 + ... + a_{2d}X_d \\ \vdots &\quad\quad \vdots \\ Y_d &= a_{d1}X_1 + a_{d2}X_2 + ... + a_{dd}X_d \end{aligned} \tag{2.1}$$

The principal components are those uncorrelated linear combinations $Y_1, Y_2, ..., Y_d$ with maximum variance. Because it preserves variance in the data rather than maximizing discrimination between classes, PCA does not guarantee good classification accuracy. Further, the method does not exploit the ordered information in the spectral bands and is not class dependent.

- Minimum Noise Fraction Transform: The minimum noise fraction (MNF) transform was developed as an alternative to PCA, which is affected by outliers. Instead of preserving the variance of the feature space, MNF maximizes the signal-to-noise ratio (SNR) by segregating additive noise that is assumed to be uncorrelated with the signal [33]. The MNF transform chooses linear transformations $Y_i = aX,\ i = 1, \ldots, d$ such that the noise fraction for Y_i is maximum among all linear transformations orthogonal to $Y_j,\ j = 1, \ldots, i-1$. The transformed data are arranged in bands of decreasing noise fraction (increasing SNR). The MNF transform is widely used in remote sensing applications. However, similar to PCA, it has no class dependent information and does not exploit band adjacency or relate to improved class discrimination.

Although feature extraction algorithms are generally computationally superior and yield higher accuracies than feature selection algorithms, they too have weaknesses. For instance, PCA and MNF transforms project the feature space linearly and cannot handle data that are embedded in a nonlinear space. They are also problematic if estimation of the covariance matrix is required, but the number of training samples is small. Further, the resulting weights on the original bands are often sensitive to variations in the data, thereby making them less desirable in the knowledge transfer framework, which is discussed later in this chapter.

Support Vector Machines

The Support Vector Machine (SVM) is a nonparametric classification method that searches for the optimal hyperplanes, defined as the linear decision function which maximizes the margin between the two classes [16].

Assume a binary classification problem with input space $\mathbf{X}$, binary class labels $Y : Y \in \{-1, 1\}$ and training samples:

$$S = (y_1, \mathbf{x}_1), ..., (y_l, \mathbf{x}_l), \quad y_i \in \{-1, 1\}. \tag{2.2}$$

The SVM seeks the optimal hyperplane

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \tag{2.3}$$

with variables $\mathbf{w}$ and b such that

$$y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \quad i = 1, ..., l. \tag{2.4}$$

The resulting minimum distance between two class groups in the new projection is

$$\rho (\mathbf{w}, b) = \frac{2}{|\mathbf{w}|} = \frac{2}{\sqrt{\mathbf{w} \cdot \mathbf{w}}} \tag{2.5}$$

For a given training set S, the values of $\mathbf{w}^*, b*$ that maximize $\rho (\mathbf{w}, b)$ are solutions to the quadratic optimization problem:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\mathbf{w} \cdot \mathbf{w} \\ s.t. \quad & y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \quad i = 1, ..., l. \end{aligned} \tag{2.6}$$

If the given training sample set is linearly separable, the optimization problem (2.6) has feasible solutions. The optimal solution $\mathbf{w}$, and b determines the hyperplane

that maximizes the margin between two different classes in the new projection. (See Figure 2.1 for an illustration of a 2-dimensional input problem.) Since it avoids the estimation of the class distributions, and the number of variables in the dual of this quadratic optimization problem does not depend on the dimensionality of the input space, SVMs often perform well when classifying problems with high dimensional input spaces [16].

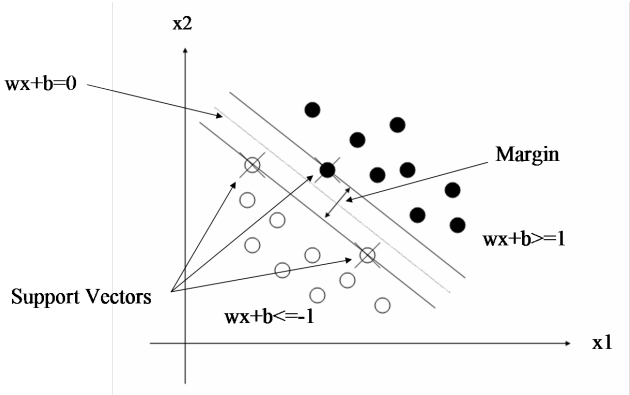

Figure 2.1: Graphical representation of SVM optimization problem

Not every problem is guaranteed to be linearly separable, so a soft margin hyperplane SVM was developed to separate the training set with a minimal number of errors [16]. The associated optimization problem introduces non-negative slack variables ξ_i and becomes

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\mathbf{w}\cdot\mathbf{w} + CF\left(\sum_{i=1}^{l}\xi_i\right) \\ s.t. \quad & y_i\left(\mathbf{w}\cdot\mathbf{x}_i + b\right) \geq 1 - \xi_i \quad i = 1, ..., l \\ & \xi \geq 0, \end{aligned} \tag{2.7}$$

where $F(u)$ is a monotonically increasing convex function, and $C > 0$ is a user-defined penalty constant. The optimization problem in (2.7) allows class samples to be misclassified, while incurring a penalty cost $CF(u)$. It has been shown that

when the size of the training sample is small, it is important to select an appropriate C to mitigate the effect of outliers in the training set.

While SVM learns linear decision boundaries, most of the real-world classification problems have nonlinear decision boundaries. SVM, therefore, maps the samples into a higher dimensional space where the classes can be separated by linear hyperplanes, thereby obtaining a nonlinear decision boundary in the original feature space. The nonlinear decision surfaces are calculated through a kernel function K, such that $K(x_i, x_j) = (\Phi(x_i) \cdot \Phi(x_j))$ where Φ maps the current space $\Re^d$ to a higher dimensional space $\Re^D$ $(d \ll D)$. Some common kernels include:

$$PolynomialKernel: \qquad K(x_i, x_j) = (x_i \cdot x_j)^d \tag{2.8}$$

$$GaussianKernel: \qquad K(x_i, x_j) = \exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2}\right) \tag{2.9}$$

$$RadialBasisFunctionKernel: \qquad K(x_i, x_j) = \exp\left(-\gamma||x_i - x_j||^2\right). \tag{2.10}$$

Although many kernels, such as polynomial and Gaussian, have been implemented for hyperspectral data classification, the highest classification accuracies have typically been achieved using the radial basis function (RBF) [11, 20]. These studies also found that RBF-SVM is very sensitive to the values of parameters, and the tuning process is tedious and slow. In addition, SVM only handles binary classification problems in its traditional implementation and requires an output space decomposition approach to extend it to the multi-class problem. One of the important issues for this study is to reduce the dependencies on the RBF kernel and reduce the time required for the tuning process. Details are contained in the next chapter.

2.1.2 High Dimensionality Output Space Problem

Land cover classification problems often have tens of outputs (class labels). However, many classification algorithms, including SVM, are binary in nature, whereby a

group of observations is separated into two groups based on the input spaces of those observations. Multi-class problems can be decomposed into simpler binary problems via a number of output space decomposition methods such as pairwise and error correcting output codes (ECOC) [24]. The pairwise decomposition method often yields good results because the classifiers can be tuned to pairs of single classes, but requires a separate classifier for each pair of classes and a combining method. For example, a 14 class problem requires 91 classifiers to cover all pairwise combinations. In the ECOC, a C-class problem (C is the total number of classes) is decomposed into $\overline{C}$ binary problems whereby the original class is then encoded into a $\overline{C}$ binary vector of a coding matrix [24]. The ECOC does not always provide good classification results because the coding matrix design ignores the natural groupings of output classes, an important characteristic of land cover classification problems [63].

Binary Hierarchical Classifier

Binary trees, which often provide an attractive approach for decomposing large output space problems, can be constructed using a variety of splitting functions involving single or multiple features and output classes. To address the high dimensional output problem while exploiting the affinity for spectrally similar classes, Kumar *et al.* proposed a binary hierarchical classifier (BHC) [49] to decompose an n-class problem into a binary hierarchy of simpler 2-class problems that can be solved using a corresponding hierarchy of classifiers, each based on a simple linear discriminant. The method was extended by Morgan *et al.* [56] for small training samples using an adaptive best-bases BHC, which exploits the class specific correlation structure between sequential bands of hyperspectral data and utilizes an adaptive regularization approach to stabilize covariance estimates. The BHC is based on two algorithms: Fisher's linear discriminant and GAMLS, a framework that decomposes (meta)-classes into finer, more homogeneous subsets. A series of studies based on the BHC

have used best-bases and polyline approximations and ensemble methods to handle the small sample size problem [56, 36].

- Fisher's linear discriminant : Fisher's linear discriminant is designed to handle binary separation by projecting the original feature space of two (meta)-classes onto the real line using Fisher's projection. Here, the original multivariate feature $\mathbf{x}$ is transformed into univariate observations y such that the $y_1', \ y_2'$ is derived from two populations π_1 and π_2. The desired linear combination maximizes the ratio of the distance between classes to the distance within classes $\widehat{y}$

$$\widehat{y} = \widehat{\mathbf{a}}\mathbf{x} = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)' \, \Sigma_{pooled}^{-1} \mathbf{x} \frac{(\overline{y}_1 - \overline{y}_2)^2}{\sigma_y^2} \tag{2.11}$$

  [26]. A novel observation is classified by projecting it to the same real line and comparing the distances to the two (meta)-classes' means. It is then designated as a member of the nearest class. Because they exclusively handle binary separation, output space decomposition methods are required to extend this classifier to handle multi-class classification problems.

- Generalized Associative Modular Learning System: GAMLS : A binary decomposition framework was recently developed that attempts to achieve high accuracy while exploiting class similarities [47]. Its main function is to group classes into two (meta)-classes based on the natural affinities among the classes so that the binary problem of separating these two (meta)-classes is relatively easy. By recursively applying this approach to the two subgroups, a binary hierarchical output space decomposition is achieved. This approach was used in the Generalized Associative Modular Learning System (GAMLS) [47], a simulated annealing-based class decomposition algorithm utilized by the Binary Hierarchical Classifier (BHC).

The hierarchical output space decomposition scheme is an NP-Hard problem. In top-down BHC, a (meta)-class Ω_n is partitioned into two (meta)-classes Ω_{2n} and Ω_{2n+1}. Each label set $L_n : L \in \Omega_n$ is associated with two (meta)-classes (A). The initial association is $A = (1, 0.5, ..., 0.5)$. If $A_i = 1$, L_i is associated with Ω_{2n}. When $A_i = 0$, L_i is arbitrarily associated with Ω_{2n+1}. The association is defined as the posterior probability $P(\Omega_\rho|L_i)$ of a class L_i belonging to a particular (meta)-class $\Omega_\rho, \rho \in \{2n, 2n+1\}$ [55]. Given A, Fishers linear discriminant $\Psi(x|A)$ is employed to separate Ω_n and Ω_{n+1}. The value of $\Psi(x|A)$ is used to calculate the log-likelihood Li of $L \in \Omega$,

$$Li(L|\Omega_\rho) = \frac{1}{N} \sum_{x \in X_L} \log p(\Psi(x|A)|\Omega_\rho), \ \rho \in \{2n, 2n+1\} \ \forall L \in \Omega. \qquad (2.12)$$

Equation (2.12) is used to update the association vector A until the difference between values of Equation (2.12) in two sequential iterations is less than a user-specified threshold. At this time, A is rounded to the nearest integer $\{1, 0\}$. Each class in the (meta)-class Ω_n is assigned to Ω_{2n} or Ω_{2n+1} [19].

Previous studies have demonstrated that this framework has the following advantages for classification of remotely sensed data with large output spaces: 1) the order of the number of binary classification problems reduces from $O(C^2)$ to $O(C)$; 2) the impact of the small sample problem is mitigated; 3) the framework provides a natural, intuitive structure that exploits affinities between classes [49].

Although GAMLS artificially increases the number of labeled samples at the top of the hierarchy and mitigates the effect of high dimensional input data, for internal nodes that are closer to the pure leaf nodes, the number of labeled samples is small while the feature dimension remains high. Algorithms based on discriminant functions require estimation of the vector-valued mean and covariance matrices $\Sigma = Cov(\mathbf{X}) = E(\mathbf{X} - \mu)(\mathbf{X} - \mu)'$ of each class label to construct a decision

boundary that separates the samples. For a d -dimension input space training sample set, where d is the dimension of $\mathbf{X}$, there are $d(d-1)/2$ parameters in the estimated covariance matrices $S = \frac{1}{d-1}\sum_{j=1}^{d}(\mathbf{X}_j - \overline{\mathbf{X}})(\mathbf{X}_j - \overline{\mathbf{X}})'$ necessitating a minimum of $d+1$ samples. In general, the literature recommends a ratio of observations to dimensionality of 4-10 times for linear classifiers. While quadratic classifiers requiring the estimation of the covariance matrix may perform better than linear classifiers, the recommended number of observations is related to the square of the dimensionality [55]. In order to reduce the impact of the effect, Morgan *et al.* [56] proposed a best-bases band aggregation approach (BB-BHC).

Similar to other feature extraction algorithms, BB-BHC requires a careful tuning process and results in loss of generalization. One feasible solution for improving BHC involves replacing its base classifier, Fisher's linear discriminant, with the more powerful SVM classifier. The hierarchical SVM (HSVM) proposed by this study is inspired by this concept. More details about the integration of the hierarchy and SVM are presented in Chapter 3.

2.1.3 Ensemble Methods

The theory and practice of classifier ensembles provides another way of alleviating sample size and high dimensionality concerns [50]. Researchers usually consider ensemble methods as a means of mitigating the impact of a limited number of labeled samples. Skruichina and Duin first proposed the implementation of ensemble methods in land cover classification [72]. They investigated the impact of bagging [8], boosting [68], and random subspace methods [37] when applied to hyperspectral data. Their study found that when the input space is large, random subspace feature selection can provide improved classifier diversity, while stabilizing parameter estimates, by randomly reducing the number of inputs to each classifier in the ensemble and constructing multiple classifiers in the resulting random input space. Because

an ensemble of classifiers is used, each classifier only utilizes a random subspace of the original feature space, and the estimated parameters are stable in each classifier. Additionally, using an ensemble of classifiers reduces the model variance and yields higher classification accuracies. The method can provide increased classifier diversity while stabilizing parameter estimates, and is effective for solving problems with redundant input features (e.g., hyperspectral data) and outliers in the training data. However, the random subspace method is computationally costly due to the large (often 50-100) number of classifiers required in the ensemble.

Recently, approaches referred to as "random forests (RF) of classifiers" have been proposed. These involve developing multiple trees from randomly sampled subspaces of input features, then combining the resulting outputs via voting or a maximum *a posteriori* rule [9]. These methods typically achieve superior generalization for small training samples, but are inherently computationally intensive. Details of the RF approach are described in Appendix B.

2.2 Contextual Classification

Although many algorithms have been developed to utilize spatial information, most methods are designed to reduce the impact of outliers in homogeneous areas or to model texture patterns, and can only be used with low dimensional multispectral data. Meanwhile, most studies in land cover classification that use hyperspectral data are pixel-based and do not incorporate contextual information. Because of the medium (30 m) spatial resolution of Hyperion data, pixels on the class boundaries are often comprised of multiple classes and pose a problem for pixel-based algorithms. To better distinguish these mixed pixels, contextual information should be incorporated into the classification process. In this section, algorithms such as a stacked vector approach, image segmentation and majority filtering are briefly introduced. The popular Markov random field (MRF)-based algorithms are presented

in more detail.

2.2.1 Common Approaches for Including Spatial Context

Previous studies that include spatial context are members of three general categories of approaches:

- A stacked vector approach [35], whereby the original averaged bands or Fourier transforms of neighboring features are concatenated with the original spectral vectors. While these approaches provide insight into the spatial neighborhood, they can be handicapped by an insufficient number of labeled samples and relatively high dimensional inputs. In addition, a stacked vector approach cannot adequately classify pixels at the image boundaries because of large changes in spectral signatures between the central target pixel and its neighbors. However, including all the members of the neighborhood in the input vector would create an exorbitantly large input space.

- Image segmentation is widely used to incorporate spatial information. There methods divide images into many homogeneous segments according to their spatial-spectral proximity using either bottom-up or top-down approaches [42, 53]. Subsequent classification is performed by comparing the similarities of labeled samples to the characteristics of each segment. These approaches produce segments that are spatially and spectrally homogeneous. Unfortunately, the classification accuracies of these algorithms are very sensitive to the initial segmentation settings. For example, the number of segments and the specifics of these segments are critical to achieving a good classified map.

- Majority Filtering is a simple, commonly used approach which performs majority voting [21] after the image is first classified by a pixel-wise classifier. The majority filtering process assigns a pixel's label according to its first-order or

second-order neighbors. If the local neighborhood is dominated by one class, the label of the targeted pixel is changed to reflect the majority. This process removes outliers in homogeneous areas, but the resulting classified maps are often blocky and do not properly identify class boundaries.

2.2.2 Markov Random Fields

For over a decade, Markov random fields (MRF) [30, 23] have been widely used for incorporating spatial-spectral information in the classification process. The general assumption of MRF is that $\pi(c)$, the prior probability of each class c, can be modeled as a discrete MRF:

$$\pi(c|c_i, \forall i \in I) = P(c|c_s, \forall s \in S). \tag{2.13}$$

I is the whole image, S is the local neighborhood, and $s \in S$ denotes pixels in the neighborhood. The isotropic behavior and the local dependencies make MRF an ideal approach for learning contextual information from S. According to the Hammersley-Clifford theorem, $\pi(c)$ is equivalent to a Gibbs distribution such that

$$\pi(c) = 1/Z \ \exp(-\sum V_s(c)). \tag{2.14}$$

Z is a normalizing constant and V_s is the energy function of the Gibbs distribution for s. Selecting the V_s function is an important issue for MRF estimation.

To determine the unknown class label c of each image pixel, spatial information is utilized by a Bayesian estimator; the maximum *a posteriori* (MAP) classifier selects the optimal $\hat{c}$, which is given by:

$$\hat{c} = \arg\min_{c} \left\{ -\log P\left(X|c\right) + V_s(c) \right\}, \tag{2.15}$$

where X is the input space, and $P\left(X|c\right)$ is the conditional probability of the input conditioned on class c. The optimization problem is a non-convex nonlinear problem

and can be solved by various heuristic approaches. The most common algorithm uses simulated annealing (SA) to select the optimal c

$$\hat{c} = \arg\min_{c} 1/Z \ \exp\left\{-\frac{1}{T}U(x)\right\}, \tag{2.16}$$

with

$$U(x) = \sum_{s \in S} \log P\left(X|c\right) + V_s(c).$$

The method converges to its optimal solution as the temperature T is slowly lowered to zero. Previous studies showed that c will converge to $\hat{c}$ almost surely, but its convergence rate is very low.

Iterated conditional modes (ICM) [5] is the most widely used approach to determine the MRF parameters. ICM starts with an initial classified image and recursively reduces the total energy until it converges to a local minimum. The goodness of the final classified image is highly dependent on the quality of the initial classified image. Both MRF using the SA heuristic and ICM have potential problems when applied to hyperspectral data. Because of the curse of dimensionality, the MAP classifier is often impacted by small training data sets, which result in near-singular covariance matrices. To overcome this problem, Jackson and Landgrebe proposed an adaptive Bayesian classifier [40] that uses a semi-supervised approach to increase the number of labeled samples. The resultant covariance matrix is more stable, but excessive computational time is required to obtain the MAP solution.

Recently, Camps-Valls *et al.* [13] proposed an algorithm that learns kernel functions of spatial and spectral similarities of hyperspectral data separately. It then combines the two kernel functions to form a kernel machine that satisfies Mercer's conditions, in which, the new kernel is also positive definite. Because of the variety of spatial textures, it is difficult to include all scenarios in one spatial kernel. The results are obtained by experimenting with different combinations of kernels. Thus,

the tuning process is typically time-consuming.

2.3 Knowledge Transfer Problems

The knowledge transfer problem is a relatively new research topic in land cover classification. Previously, researchers assumed that the spectral distributions of the training data set and test data set were the same. In reality, spectral signatures change with many uncontrollable factors, and it is always difficult to obtain enough ground truth pixels for a new location. Therefore, having a more robust classification algorithm is critical to addressing the knowledge transfer problem. The semi-supervised approach proposed in [39, 71] augments the training set with unlabeled data to increase the number of labeled samples, but cannot handle changes in spectral signatures.

In this section, the characteristics of ensemble methods outlined in Section 2.1.3 that are relevant to knowledge transfer problems are discussed. The idea of active learning is also discussed. Related studies that use these approaches in land cover classification are reviewed.

2.3.1 Ensemble Methods

Because ensemble methods use randomization and sampling approaches, they are able to cover multiple scenarios, which are typically encountered in knowledge transfer problems. Thus, the overall accuracy is typically improved when ensemble methods are applied to the areas that have slightly different spectral signatures. The RF approach developed in parallel to this study by the UT Remote Sensing Group and presented in Appendix B assumes that no information is available to users from new regions. In contrast to many algorithms that are able to adapt changes, ensemble methods succeed because of their diversity (see Appendix B).

Rajan *et al.* [64] follow a different assumption that very limited numbers

of labeled samples can be obtained for updating the original model. The method extends our Ham *et al.* paper [34], which achieves diversity of an ensemble by creating a large set of classifiers, by randomly switching the internal (cousin) nodes of the BHC structure at a given level of the tree. Best bases band aggregation is employed for feature reduction. After these models are created, they use the given limited new labeled samples from the new areas to evaluate the models and determine the weights of the individual classifiers for a weighted ensemble. The combination of ensemble approaches and the small portion of labeled data from the new areas produced very encouraging results.

2.3.2 Active Learning

Both semi-supervised classification and active learning use unlabeled samples to update classifier decision boundaries. The difference between these two approaches is that active learning uses a model that selects unlabeled samples intelligently for subsequent identification of the label, so the new decision boundary moves with these newly classified samples, while semi-supervised learning selects samples randomly.

Active learning has been widely used in situations where labeled samples are difficult or costly to obtain. For its application in land cover classification, Rajan *et al.* [65] implement active learning to help solve the knowledge transfer problem. They propose a new active learning technique that can be used in conjunction with any classifier that determines the decision boundary via (an estimate of) *a posteriori* class probabilities, i.e., classifiers that are probabilistic or generative rather than discriminative. The active learning process is guided by the *a posteriori* probability distribution function $P(Y|\mathbf{X})$ which is the probability of a sample belongs to class Y if it has an input data $\mathbf{X}$. The goal is to increase the information gain between $P_{D+L}(Y|\mathbf{X})$ and $P_{DL}(Y|\mathbf{X})$, the *a posteriori* probability density functions estimated from $D+L$ and DL, where D is the original data set and L represents the new area.

Maximizing the expected information gain between $P_{D+L}(Y|X)$ and $P_{DL}(Y|\mathbf{X})$ is equivalent to selecting the data point $\mathbf{x}$ from $D \cup L$ such that the expected Kullback-Liebler divergence between $P_{D+L}(Y|\mathbf{X})$ and $P_{DL}(Y|\mathbf{X})$ is maximized. That is, those data points that change the current belief in the posterior probability distribution the most are selected. This intelligent sample selection process helps by guiding the new classifier, which adapts to changes of spectral signatures from one location to other locations.

Most of the algorithms discussed in this chapter that perform land cover classification using hyperspectral data require a long and careful tuning process. In the next three chapters, a series of new classification processes that have reduced computation times and reduce the knowledge threshold required to utilize hyperspectral data are presented.

Chapter 3

Integrating Support Vector Machines in a Hierarchical Output Space Decomposition Framework

Support vector machines (SVM) [16], introduced in Section 2.1.1, have gained attention in the remote sensing community because of their ability to accurately classify high dimensional data using a small number of labeled samples. Since SVM does not involve parameter estimation and is relatively unaffected by the high dimensionality of limited training data, it is ideal for hyperspectral data classification [27]. Results of implementing SVM for hyperspectral data classification presented by Camps-Valls *et al.* [12] illustrate the high classification accuracies achieved by SVM. In this study, the weight of each band as determined by the SVM was also used to rank the bands. In addition, Melgani and Bruzzone [54] found SVM to be superior to the RBF neural network and k-nearest neighbor when used to classify

hyperspectral data.

The SVM is inherently designed for binary classification problems. Traditional class decomposition approaches have been investigated for extending the SVM approach to handle multi-class problems [38, 54]. One-vs-all and the ECOC [24] decomposition methods can sometimes achieve high classification accuracies using the associated class groups, but often require a complex SVM kernel to construct the decision boundary [62]. This results in a time consuming, tedious parameter tuning process. The goal of this chapter is to present a class decomposition algorithm integrated with the support vector machine framework. The new algorithm handles large input and output space problems and requires only a limited amount of tuning, but achieves high classification accuracies. This integrated hierarchical support vector machine is the first contribution of this research.

3.1 Methodology

The generalized associative modular learning system (GAMLS), a hierarchical class decomposition algorithm that is integrated with the Fisher's linear discriminant (FLD) for classification, was presented in Section 2.1.2. The BHC has the following advantages:

1. The order of the number of binary classification problems is reduced from $O(C^2)$ to $O(C)$, while C is the number of classes.

2. The impact of the small sample problem is mitigated.

3. The framework provides a natural, intuitive structure.

Since SVM and FLD are both binary, linear discriminant classifiers, an intuitive approach is to build the binary hierarchical tree using GAMLS and replace FLD with SVM [20]. A new hierarchical class decomposition algorithm which uses max-cut decomposition to fully integrate the margin maximization approach used by

SVM is proposed. The two components of the proposed method are explained in the following section.

3.1.1 Binary Hierarchical Support Vector Machines using GAMLS

This preliminary work explored the possibility of developing a binary hierarchical support vector machine (BH-SVM) classifier to handle hyperspectral data. The SVM classifier (BH-SVM) was investigated both as a means of dealing directly with high dimensional input data and providing an alternative to the weak Fisher's linear discriminant classifier. SVM classifiers were incorporated into the BHC during a second stage after the hierarchy had been developed using the annealing-based GAMLS [20, 62]. The hierarchical class decomposition and SVM binary classifiers for this BH-SVM method were only integrated in the Fisher projection framework. Overall classification accuracies obtained by the BH-SVM are high, but it requires a time consuming, and careful tuning process.

3.1.2 Hierarchical Support Vector Machines using the Max-cut Algorithm

To fully integrate SVM into the hierarchical decomposition structure, a new max-cut decomposition method is proposed. It provides an alternative to GAMLS for obtaining the hierarchical class decomposition by searching for the best binary partitions separated by the maximum total distance and allowing the natural incorporation of the SVM classifier. The proposed HSVM method is based on a max-cut hierarchical output space decomposition algorithm and uses SVM as the base classifier at each internal node. Because both SVM and max-cut decomposition utilize the pairwise distances between samples of different classes, HSVM is considered to be an "integrated" algorithm. The max-cut problem is first presented as background, then details of the HSVM are provided.

Max-Cut Problem

The max-cut problem is a combinational optimization problem whereby an undirected graph with nonnegative edge weights is partitioned into two groups such that the cut between these two groups has the maximum weight [76]. Define an undirected graph $G = (N, E)$ where N represents nodes, E represents edges of the graph, and $w_{ij} \geq 0$ represents the weight of an edge linking nodes i and j. The objective is to find the best binary partition that has the cut $\delta(K^*)$, $K^* \subseteq N$, and $\{ij \in E : \ i \in K^*, j \notin K^*\}$ that has the maximum weight:

$$w(\delta(K^*)) = \sum_{ij \in \delta(K^*)} w_{ij}. \tag{3.1}$$

The graph is assumed to be complete by setting $w_{ij} = 0$ for all non-edges ij. An example of this max-cut problem is illustrated in Figure 3.1. The cut (edges) between partition A and B is maximized.

The max-cut problem can be represented using an integer quadratic programming formulation with decision variables $x_i \in \{1, -1\}, \ \forall i \in N$. For a cut $\delta(K)$, $x_i = 1 \iff i \in K$. If $ij \in \delta(K)$, $x_i x_j = -1$. Thus:

$$w(\delta(K)) = \frac{1}{2} \sum_{i<j} w_{ij} \left(1 - x_i x_j\right) \tag{3.2}$$

and the resulting max-cut integer quadratic problem is:

$$\begin{aligned} \max \quad & w(\delta(K)) \\ s.t. \quad & x_i \in \{+1, -1\}, i \in N. \end{aligned} \tag{3.3}$$

The max-cut problem is known to be NP-hard [60], the combination of feasible solutions grows exponentially with N. Equation (3.3) can be relaxed and

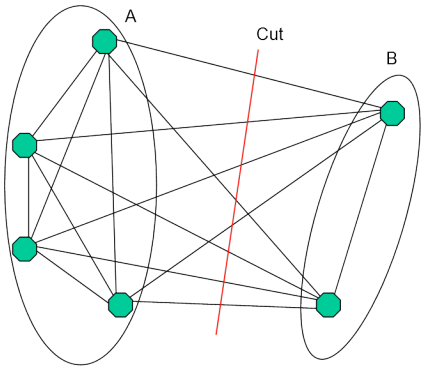

Figure 3.1: Max-Cut Example

formulated as in Equation (3.4). The dual of this relaxed max-cut problem can be formulated as a constrained quadratic problem and solved using semi-definite programming (SDP).

$$\begin{aligned} \max \quad & w(\delta(K)) \\ s.t. \quad & x_i^2 = 1, i \in N. \end{aligned} \tag{3.4}$$

The dual of max-cut problem provides an upper bound of the max-cut problem. An extension of an interior point method [57] provides a computationally efficient method for solving the semi-definite problem. The relaxed max-cut problem solved using SDP can achieve near optimal results [31].

Hierarchical Support Vector Machine

To exploit natural class groupings in combination with the SVM classifier, the proposed max-cut hierarchical output space decomposition method searches for the maximum total distance between two class partitions. The original class samples are treated as an undirected graph G where node n_i represents class i and the

non-negative weight:

$$w_{ij} = \frac{1}{2} \sum_{\forall x} \left(f_i(x) \log \frac{f_i(x)}{f_j(x)} + f_j(x) \log \frac{f_j(x)}{f_i(x)} \right) \tag{3.5}$$

is the average Kullback-Leibler distance [45] between the density function of class i and class j. The new HSVM approach solves this max-cut problem to achieve the required unsupervised class decomposition at each node of the binary hierarchical structure. As with the original BHC, the output space is hierarchically decomposed into pure leaf nodes that have only one class label at each node (see Figure 3.2). Since this max-cut unsupervised decomposition uses total pairwise distance measures to investigate natural class groupings, the hierarchical structure results in a fast and intuitive SVM training process that requires little tuning. As demonstrated in the following experiments, the method also has both high accuracy levels and good generalization.

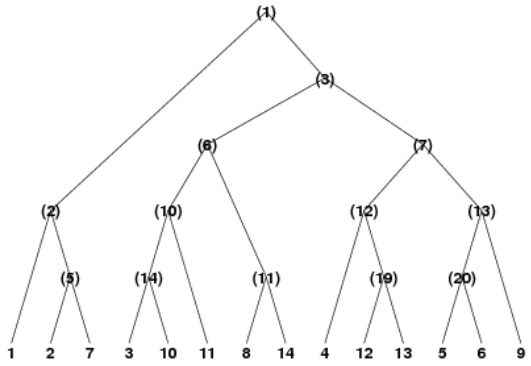

Figure 3.2: Typical HSVM hierarchical structure

3.2 Experiments

The new HSVM was applied to Hyperion hyperspectral data collected over the Okavango Delta of Botswana and data acquired by NASA Airborne Visible/Infrared

Imaging Spectrometer (AVIRIS) over the Kennedy Space Center, Florida. The hierarchies created by GAMLS and the max-cut algorithm were compared. Each algorithm was evaluated according to its structure and the complexity of its decision boundaries via the average number of support vectors. Second, HSVM was applied to the two study sites. The classification accuracies and generalization capability obtained by HSVM were compared to those achieved by the best basis hierarchical classifier presented in Section 2.1.2 and the binary hierarchical support vector machine [20] with linear (BH-SVM(L)) and RBF kernels (BH-SVM(R)). The processing times required by the classifiers are compared.

3.2.1 Comparing GAMLS and Max-cut Hierarchy

The goal of this section is to compare various aspects of the max-cut and GAMLS based SVM hierarchical classification approaches. Examples shown in Figure 3.3 and Figure 3.4 illustrate the importance of having a good decomposition algorithm. Figure 3.3 shows an ineffective hierarchy that requires complex decision boundaries to separate classes 1, 2 and 3. Figure 3.4 demonstrates that this same problem can be separated linearly if Class 3 is split at the top node of this hierarchy.

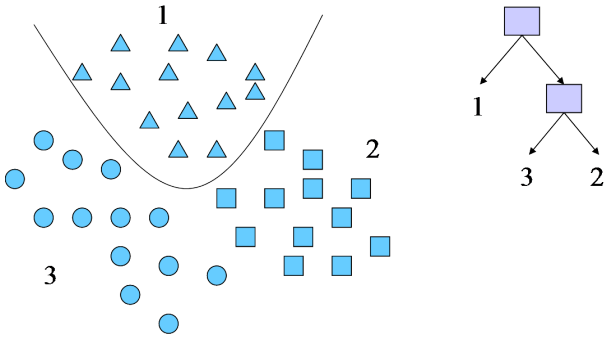

Figure 3.3: Complex Decision Boundaries

The linear decision boundary obtained by SVM is controlled by a limited number of samples called support vectors (SVs). A complex decision boundary

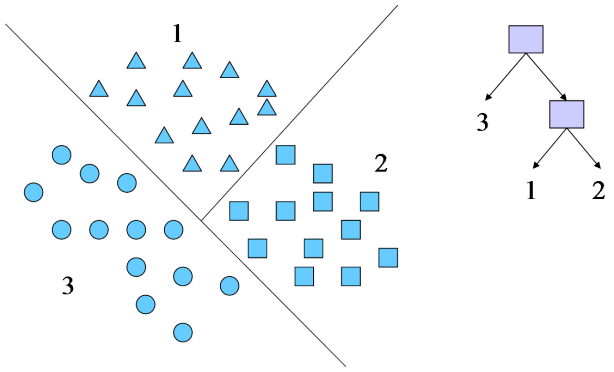

Figure 3.4: Linear Decision Boundaries

usually requires a larger number of SVs to solve the binary classification problem. This study uses the average number of SVs required as a quantitative measure to evaluate hierarchies created by GAMLS and the max-cut approach. Figure 3.5 shows that the average number of SVs required by the max-cut decomposition for various sampling rates which is defined in Appendix A.1 is consistently lower than that required when the class decomposition is GAMLS based, which indicates that the hierarchy created by the max-cut approach supports a good decomposition that requires less tuning and has a faster training/testing process.

3.2.2 Classification - Original Training and Test Areas

As described previously, the HSVM was applied to two research sites: Kennedy Space Center (KSC) and Okavango Delta of Botswana. Detailed descriptions of these two sites are contained in Appendix A.1 and A.2.1. The standard approach used throughout this research is described as follows: for both datasets, ten randomly sampled partitions of the training data were sub-sampled such that 75% of the original data were used for training and 25% for testing. In order to investigate the impact of the quantity of training data on classifier performance, training data were then sub-sampled to obtain ten samples comprised of 50%, 30%, and 15% of

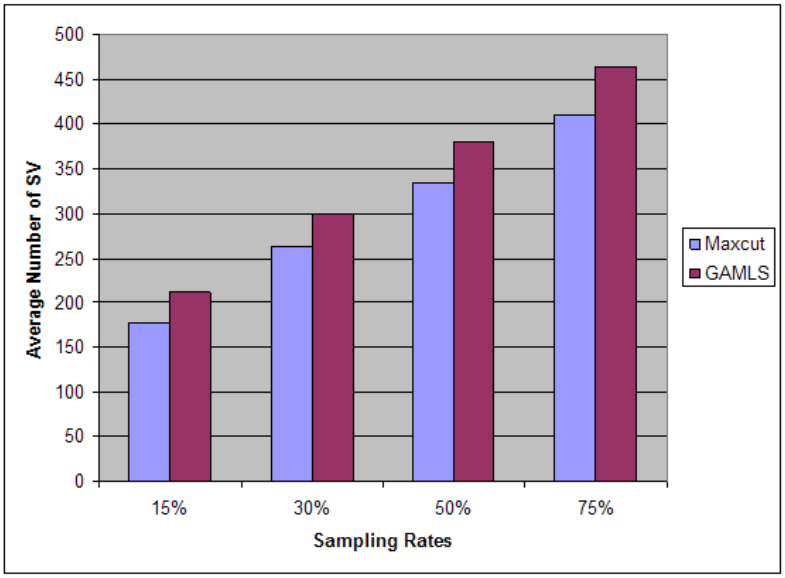

Figure 3.5: Average Number of Support Vectors Required for GAMLS and HSVM

the original training data. All classifiers were evaluated using the ten test samples composed of 25% of the original training data. A classified image created by HSVM is shown in Figure 3.6.

Comparing BB-BHC, BH-SVM(L) and HSVM(L)

Average classification accuracies of the BB-BHC, BH-SVM(L) and HSVM(L) methods for the 10 experiments conducted with each classifier using the Botswana test data are listed in Table 3.1. The table shows that classification accuracies increase while the standard deviation of accuracies decreases as the training sample size increases for all three classifiers. Average accuracies for the BB-BHC, BH-SVM(L) and HSVM are comparable at a 15% sampling rate, with BB-BHC having the lowest standard deviation. Although the standard deviation of the accuracies of the results obtained from HSVM is higher than BB-BHC at a 15% sampling rate, it is the lowest among three classifiers for all other sampling rates. This table also shows that

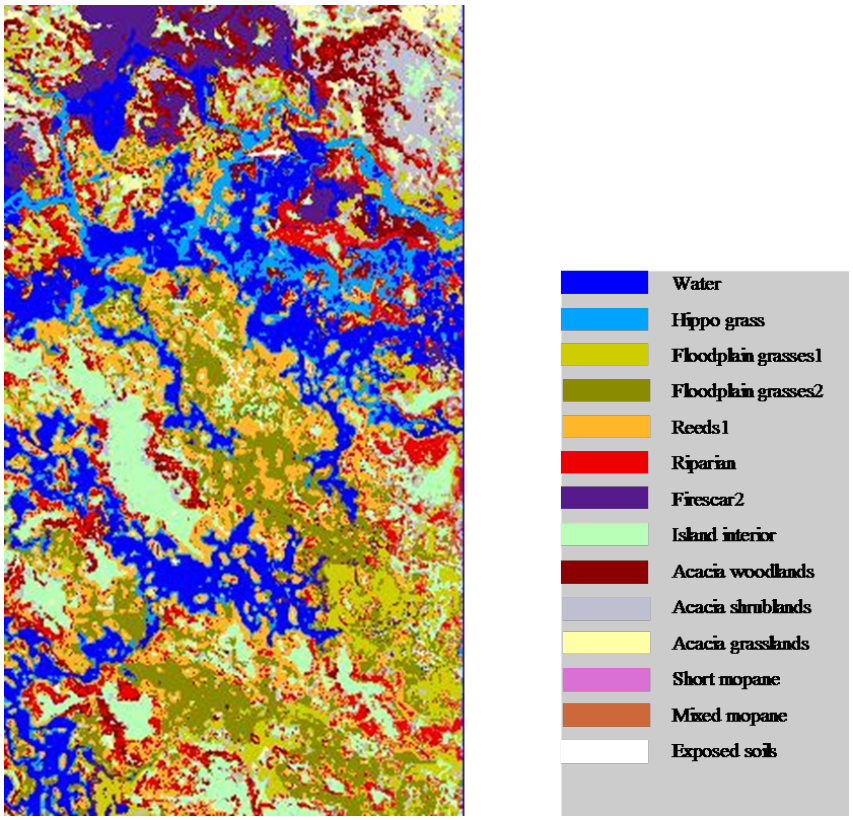

Figure 3.6: Botswana Classification Image from HSVM

BB-BHC is very competitive under the training/test setting at all four sampling rates. The accuracy differences between BH-SVM(L) and HSVM(L) demonstrate the benefits of using max-cut decomposition compared to GAMLS in the hierarchy integration. Unlike HSVM(L) that has the highest accuracy at a 75% sampling rate, the classification accuracy of BH-SVM(L), which also uses a linear kernel SVM, does not improve as the sampling rate increases from 50% to 75%. This likely indicates that a more complex decision boundary is required for this hierarchy. The importance of having a good class decomposition strategy which leads to simplification in the tuning process is demonstrated.

A similar pattern is observed in the results from the KSC AVIRIS data. For

Table 3.1: Botswana Test Data: Ave. Accuracy (Std. Dev.)

Training	BB-BHC	BH-SVM(L)	HSVM(L)
15%	89.9(1.36)	**90.7(1.38)**	90.7(2.49)
30%	91.8(1.75)	92.8(1.88)	**93.2(1.07)**
50%	92.9(0.73)	93.6(1.16)	**94.1(0.97)**
75%	94.0(0.69)	93.8(0.96)	**95.1(0.63)**

Table 3.2: KSC Test Data: Ave. Accuracy (Std. Dev.)

Training	BB-BHC	BH-SVM(L)	HSVM(L)
15%	88.9(0.70)	91.0(0.97)	**91.7(0.93)**
30%	91.1(1.43)	92.2(0.87)	**92.8(1.25)**
50%	92.4(0.97)	92.1(0.78)	**92.9(0.53)**
75%	92.9(1.01)	92.3(0.68)	**93.9(0.45)**

these data, Table 3.2 indicates that SVM classifiers yield higher accuracies than BB-BHC when the number of samples is small. As with the Botswana experiments, the performance of BH-SVM(L) did not improve further as the sampling rate increased from 50% to 75%. These results seem to indicate that the hierarchical structure created by GAMLS does not provide a decomposition framework whose decision boundaries can be subsequently improved substantially by the more powerful SVM classifier. This justifies the need for an approach which jointly exploits the power of SVM in performing the decomposition and determining the decision boundaries. The max-cut decomposition and the SVM classifier accomplish this, working as an integrated algorithm which yields higher classification accuracies.

Comparing BH-SVM(R) and HSVM(L)

As noted in Section 2.1.1, the SVM classifier can be implemented with a wide variety of kernels (e.g., polynomial, Gaussian and RBF). The most popular of these, the RBF kernel, was implemented to investigate the impact of using a more complex SVM kernel. HSVM(L) is compared to BH-SVM using the RBF kernel: BH-SVM(R)

in Tables 3.3 and 3.4. Figures 3.7 and 3.8 show the average processing time required by these classifiers.

Table 3.3: Botswana Test Data: Ave. Accuracy (Std. Dev.)

Training	BH-SVM(R)	HSVM(L)
15%	**92.3(1.15)**	90.7(2.49)
30%	**93.8(2.23)**	93.2(1.07)
50%	**96.2(0.75)**	94.1(0.97)
75%	**96.6(0.95)**	95.1(0.63)

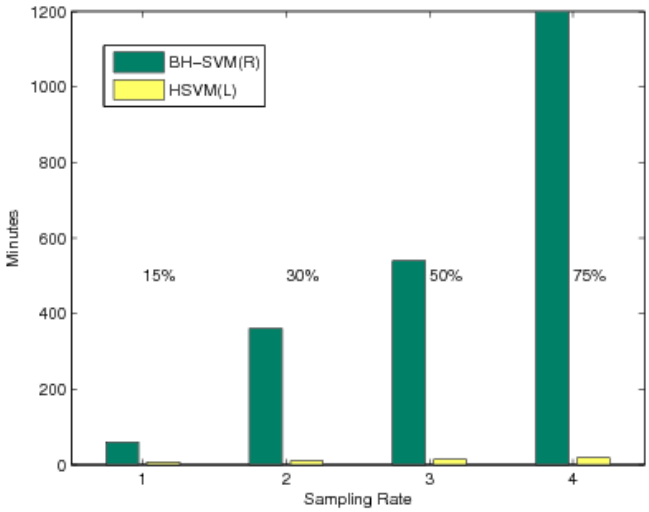

Figure 3.7: Botswana Test Data: Processing Time Statistics Using SVM Classifiers

While the BH-SVM(R) typically performed the best in terms of both average accuracies and corresponding standard deviations, the HSVM has comparable classification accuracies, sometimes at a cost of slightly higher standard deviations of these overall accuracies. Training SVM with the RBF kernel required much more time than training a linear kernel SVM. Although BH-SVM(R) achieves slightly better accuracies than HSVM(L), Figure 3.7 and 3.8 show that training time for the RBF kernel BH-SVM(R) can be as much as 100 times longer than the time required

Table 3.4: KSC Test Data: Ave. Accuracy (Std. Dev.)

Training	BH-SVM(R)	HSVM(L)
15%	91.4(1.51)	**91.7(0.93)**
30%	**93.8(0.55)**	92.8(1.25)
50%	**93.9(0.55)**	92.9(0.53)
75%	**94.6(0.68)**	93.9(0.45)

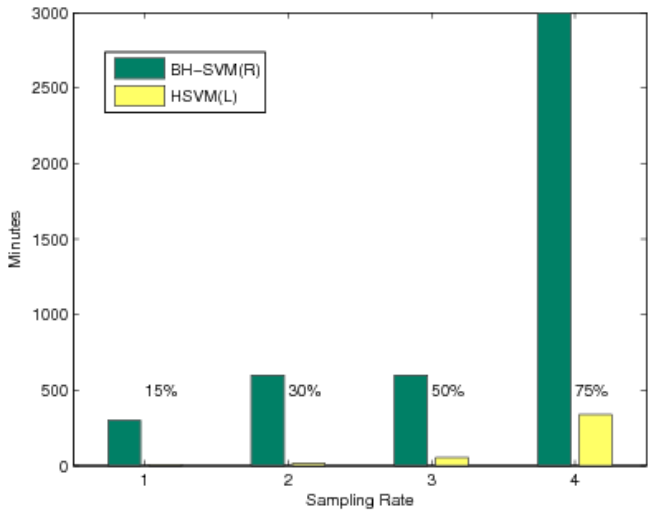

Figure 3.8: KSC Test Data: Processing Time Statistics Using SVM Classifiers

by HSVM(L) for both data sets. As the total number of training samples increases, the processing time also increases significantly for the two SVM base classifiers in these two data sets. The increase is due to the increased complexity of the dual of the quadratic optimization formulation for SVM. This situation is especially significant for the KSC data, as many training samples are available. These results assure us that HSVM dramatically reduces the tuning process, but still provides competitive classification accuracies that are nearly as high as those provided by a complex RBF-kernel BH-SVM(R).

3.2.3 Classification - Spatially Disjoint Areas

Traditionally, the training and test data are obtained as random samples of spatially co-located ground truth data, and can thus be assumed to be samples from the same distribution. In practice, it is also useful to investigate classifier performance in disjoint areas where the signatures may be somewhat different, in order to determine how much additional data labeling and retraining are required to make the model applicable to much larger areas. With this goal in mind, a "spatially disjoint" (SD) test set was also acquired from a geographically separate location at the Botswana site and used to evaluate the classifiers. Similar to the training and test data, details of the 14 classes of the Botswana SD test set are presented in Appendix A.2.1.

Overall classification accuracies and associated standard deviations for the spatially disjoint test set are presented in Table 3.5. The results show that while

Table 3.5: Botswana Spatially Disjoint Test Data: Ave. Accuracy (Std. Dev.)

Training	BB-BHC	BH-SVM(L)	HSVM(L)
15%	66.1(3.07)	66.6(3.26)	**69.3(5.06)**
30%	63.8(1.87)	64.2(3.25)	**70.4(2.17)**
50%	63.4(2.33)	63.9(2.14)	**70.8(1.32)**
75%	63.7(1.33)	62.5(1.30)	**70.3(0.97)**

both BB-BHC and BH-SVM(L) perform well on the test samples, they do not generalize well to the new area. The average classification accuracies are the same or decrease as the number of labeled samples increases. The standard deviations of the accuracies do not follow a consistent pattern for the BB-BHC, indicating that as more labeled samples become available for training, the decision rules of these two classifiers become stronger. When they are applied to a new area for which there is no prior knowledge, such strong decision boundaries are a liability. In contrast, HSVM produced the best classification results on the SD test set. The standard deviation of HSVM also decreases as the percentage of training data increases. This

indicates that the new approach with the linear kernel is robust in adapting to changes.

Experiments presented here were obtained by using MATLAB on a 3GHz Pentium 4 CPU running Red Hat Linux system. All the processing times were recorded in minutes.

3.3 Summary

The goal of this part of the dissertation was to develop and evaluate an integrated SVM classifier that has high accuracies and is friendly in terms of training for inexperienced users. A hierarchical decomposition algorithm using SVMs to classify multi-class land cover data is proposed and referred to as the hierarchical support vector machines (HSVM). It solves a series of max-cut binary set partitioning problems to hierarchically and recursively partition the set of classes into two subsets until pure leaf nodes are obtained. SVMs are used at each internal node of the hierarchy to construct the binary decision boundary.

The new HSVM was applied to hyperspectral data acquired over the Okavango Delta of Botswana and the Kennedy Space Center, Florida. Classification accuracies and generalization capability are compared to those achieved by other competitive classification algorithms. These experiments show that the new HSVM method achieves both high classification accuracy and good generalization when sample sizes are small relative to the dimension of the input space and the output space is large. Additionally, the required processing time is impressively low. This is because HSVM uses pairwise distance measures to exploit the optimal natural class groupings that require less complex decision boundaries.

With the achievement of solving the limited number of labeled samples described in Chapter 1, this proposed HSVM classifier is also implemented in the next chapter to work with a stacked vector approach to learn the contextual information

in improving the classification accuracies.

Chapter 4

Knowledge-Based Spectral Stacking for Spatial-Spectral Classification

Hyperspectral data provide new capabilities for discriminating spectrally similar classes. However, the tails of class signatures often overlap in multiple narrow bands. Additionally, many hyperspectral sensors, including Hyperion, have a push-broom design and are challenging to calibrate. Spatial neighborhood information, which is often more reliable but difficult to analyze, provides an alternative source of information which should be utilized in conjunction with spectral data to label the class pixels. Unfortunately, some traditional approaches introduced in Section 2.2 suffer from computational requirements and curse of dimensionality issues when applied to hyperspectral data. The second research focus of this dissertation is to develop a viable approach for incorporating spatial information in the analysis of hyperspectral data acquired over natural scenes, which often exhibit complex boundaries between classes. Section 4.1 provides an introduction to this spatial-spectral classification algorithm. The new proposed method of knowledge-based

spectral stacking is introduced in Section 4.2.

4.1 Overview

Although many methods have been developed to utilize spatial information in classification, most are designed to reduce the impact of outliers in homogeneous areas or to model texture patterns and are practical only for low dimensional inputs such as multispectral or multi-frequency SAR data. Most studies in land cover classification that use hyperspectral data are pixel-based. In addition to offsetting calibration related issues, spatial information is also useful for improving classification accuracy for medium resolution space-based hyperspectral data. For example, at 30m spatial resolution, pixels on the class boundaries of Hyperion data often belong to multiple classes posing a potential problem for pixel-based algorithms. To better distinguish these mixed pixels, contextual information must be incorporated into the classification process.

This study supports the view that data from the homogeneous areas and data from mixed areas should be treated differently. In the next section, an integrated, supervised classifier that merges selected spatial and spectral information to train a hierarchical support vector machine classifier [14] is described. The new algorithm is applied to hyperspectral data collected by the Hyperion sensor on the EO-1 satellite over the Okavango Delta of Botswana. Classification accuracies and the resultant classified image are compared to those achieved by a pixel-wise classifier, majority filtering and an MRF based model developed using the iterated conditional modes approach (MRF-ICM).

4.2 Methodology

Most methods which incorporate spatial information attempt to solve this relatively complex contextual classification problem using a single approach applied uniformly throughout the image. In contrast, experiments in this study show improvements in classification accuracies and reduction in processing time by breaking the problem into two smaller problems which have different objectives and are easier to solve individually. Initial results indicated that stacking average bands of a pixel's neighbors onto the vector of the original bands improves the classification accuracy of some class samples (trees and grassland), but decreases the accuracy of others (water-related areas). This is due to the complexity of the local neighborhood. This study observed experimentally that samples of wetland classes tend to have more complex neighborhoods, while flat areas with sparse woody vegetation tend to be more homogeneous in the Botswana scenes. This indicates that it would be beneficial to first distinguish these two types of neighborhoods, and then provide associated input data that support both spectral and spatial information in an appropriate way.

The significant difference in the spatial characteristics of homogeneous and mixed neighborhoods indicates that these two types of data should be treated differently in classification. In order to handle these two types of data in a straightforward, computationally efficient way, a pre-processing approach that determines the dissimilarity among pixels in a neighborhood and separates mixed neighborhoods into similar subsets is proposed. This requires an algorithm that is capable of recognizing the natural boundaries between different class labels in the user defined neighborhood. This is accomplished by defining the dissimilarity between pixels according to their Kullback-Leibler (KL) divergence [45]. Each neighborhood is separated into more homogeneous subsets by maximizing the overall dissimilarity between these subsets; the problem is then solved by graph optimization. Max-cut splitting was

used as the preprocessing algorithm to extract spatial information. The knowledge-based stacking approach used to incorporate spatial information is outlined, and integration of the approach with the original HSVM method is described.

4.2.1 Local Binary Split Approach

Max-cut optimization, which is also used in the integrated hierarchical SVM method of Chapter 3, is applied to explore the homogeneity of the neighborhood. It is a combinational optimization problem whereby an undirected graph with nonnegative edge weights is partitioned into two groups such that the sum of the weights on edges between these two groups is maximum. (This approach was discussed in Section 3.1.2.) To evaluate the effectiveness of the proposed max-cut approach, the binary split algorithm was tested locally in several areas of the image that are either homogeneous (Figure 4.1) or mixed (Figure 4.2 and 4.3).

Figure 4.1: Island Interior: (Left) Original Image, (Right) Max-cut Result

Figure 4.2: Firescar: (Left) Original Image, (Right) Max-cut Result

These figures show that max-cut can successfully detect the class boundaries

Figure 4.3: Primary Floodplain: (Left) Original Image, (Right) Max-cut Result

by maximizing the total distance between two subsets in the defined second-order neighborhood. These images also illustrate the complexity of local neighborhoods.

4.2.2 Knowledge-based Stacking

A stacked vector approach (See Section 2.2) that incorporates the average spectral values of its neighbors improves the classification accuracy of homogeneous areas, but reduces the accuracy of the mixed areas. The goal of knowledge-based stacking is to obtain the relevant subset of homogeneous neighborhood spectral information to support the classifier. The new method developed in this study utilizes the class boundaries identified by max-cut optimization to determine locally appropriate spectral information within a second order neighborhood and incorporates the results in classifier training. The logic is illustrated by the following two scenarios:

- If a pixel is located in a homogeneous area, the average of the bands of its neighbors provides additional unbiased information with lower variability, thereby leveraging spatial smoothing.

- If a pixel's spectral signature is different from its neighbors', using the average bands of all neighbors potentially reduces the classification accuracy. Here, the averaged bands of neighboring pixels that are similar to the central target pixel provide relevant information to support the selection of a class which is dominant in the mixed pixel.

In both scenarios, the number of bands/features is doubled. At this preprocessing stage, data that increases the likelihood of a pixel belonging to the correct class is incorporated by utilizing bands of neighbors that are similar to the central pixel which is being classified. This approach mitigates the impact of spectral values that lie in the tail of the distribution of a class and tends to classify correctly near boundaries which potentially contain mixed pixels of two classes. Because it is capable of splitting neighboring pixels into more homogeneous subsets, the max-cut optimization is first applied to the whole image to find suitable neighboring pixels for knowledge-based stacking.

The increased number of features associated with this stacking approach makes a support vector machine especially desirable as a classifier for this problem. In Chapter 3, a hierarchical support vector (HSVM) was developed to handle problems that involve high dimensional inputs and complex land cover data that are difficult to discriminate. The HSVM is similarly advantageous for the spatial setting because it splits a complex multi-class problem into smaller binary classification problems.

4.3 Experiments

Since the KSC data had many base classes which were actually mixed classes, and it was impossible to extract the needed neighborhood information required for training the algorithm, the proposed knowledge stacking approach was only applied to the Botswana data collected in May of 2001. This dataset is slightly different from the fourteen class dataset that was used in Chapter 3. Spectrally overlapping classes such as floodplain grass 1 and 2 in Table A.2 are aggregated as primary floodplain in this dataset. More details of this dataset are available in Appendix A.2.2. The samples located on the class boundaries are referred to as the "extended edge" test data set, which is discussed in Appendix A.2.3.

Experiments were performed using the pixel-wise hierarchical SVM (HSVM) proposed in Chapter 3, stacked vector approach (SVA), majority filtering (MF), iterated conditional modes (MRF-ICM) presented in Section 2.2, and the proposed max-cut stacking HSVM (MC-HSVM). Average test data classification accuracies for the 10 experiments conducted with each classifier are listed in Table 4.1. Classification accuracies on the extended edge test set are presented in Table 4.2. More detailed individual class accuracies are shown in Table 4.3. Detailed pairwise comparisons are presented in the next three sections to demonstrate the performance gain due to stacking vectors using the new approach.

Table 4.1: Botswana Test Data: Ave. Accuracy (Std. Dev.)

Training %	HSVM	SVA	MF	MRF-ICM	MC-HSVM
15%	96.5(0.95)	98.4(0.99)	**98.6(0.57)**	97.5(0.71)	97.5(0.77)
30%	97.3(1.14)	98.6(0.56)	**99.3(0.11)**	98.5(0.30)	98.7(0.63)
50%	97.9(0.51)	99.1(0.26)	**99.5(0.23)**	98.9(0.12)	98.7(0.58)
75%	97.7(0.52)	99.2(0.18)	**99.7(0.13)**	99.3(0.10)	99.2(0.44)

Table 4.2: Botswana Edge Test Data: Ave. Accuracy (Std. Dev.)

Training %	HSVM	SVA	MF	MRF-ICM	MC-HSVM
15%	83.6(2.58)	79.9(2.48)	84.5(2.41)	84.5(2.47)	**86.3(2.31)**
30%	87.4(1.53)	80.9(1.12)	89.4(2.14)	88.2(1.59)	**89.9(1.49)**
50%	87.9(1.24)	81.3(0.85)	89.8(1.25)	89.3(1.12)	**90.8(1.23)**
75%	88.5(0.86)	81.4(0.48)	90.9(0.94)	89.8(0.89)	**91.9(0.66)**

4.3.1 Comparing HSVM and MC-HSVM

In both Tables 4.1 and 4.2, MC-HSVM is shown to be the clear winner over pixel-wise HSVM in terms of classification accuracies. Since both classifiers follow the same HSVM framework, the improvement in classification accuracy is unmistakably due to the knowledge-based stacking. In this pairwise comparison, MC-HSVM not

Table 4.3: Botswana Edge Test Data: Individual Class Ave. Accuracy (Std. Dev.)

Class	HSVM	SVA	MF	MRF-ICM	MC-HSVM
Water	99.2(0.54)	**100(0)**	96.0(2.43)	99.2(0.22)	98.6(0.60)
Floodplain	91.9(3.62)	70.0(0.51)	**94.3(3.56)**	93.8(2.56)	91.3(3.27)
Riparian	70.1(6.92)	75.9(1.39)	77.3(7.23)	72.2(7.81)	**80.0(3.05)**
Firescar	95.7(2.45)	78.3(5.27)	**98.8(0.37)**	98.2(1.49)	94.1(4.16)
Island Interior	88.8(5.26)	82.3(5.29)	**97.4(2.62)**	94.0(2.23)	86.7(9.58)
Woodlands	90.5(1.46)	83.0(2.31)	**98.5(0.35)**	93.1(2.37)	95.0(2.49)
Savanna	89.0(2.86)	78.8(2.53)	**99.1(0.58)**	91.5(2.55)	97.6(1.17)
Short Mopane	79.4(6.47)	**90.0(0.74)**	82.4(5.33)	79.2(4.73)	80.3(4.14)
Exposed Soils	86.5(2.04)	84.8(4.13)	66.1(5.59)	85.9(2.31)	**93.2(3.62)**

only achieves higher overall accuracies, it also provides consistently better results for individual classes in both relative homogeneous, but spectrally diverse classes (woodlands and savannah), which are small (exposed soils) or have complex geometric boundaries (riparian) with other classes. (See Table 4.3.)

4.3.2 Comparing Classification Results from SVA and MC-HSVM

Classification accuracies of simple stacked vector approach that stacks average spectral data of a second order neighborhood are presented and compared to those obtained by the proposed MC-HSVM. As discussed in Section 2.2, SVM algorithms do not filter out any redundant information and cannot handle complex contextual information. Tables 4.2 and 4.3 support this finding and show that MC-HSVM provides consistently higher classification accuracies than SVM. Since both methods use similar stacked vector approaches and the same HSVM framework, it is obvious that MC-HSVM benefits from using max-cut algorithm, which is used as an intelligent filtering process in this chapter.

4.3.3 Comparing Classification Results from MF and MC-HSVM

Although the overall accuracy tables indicate that majority filtering performs well in terms of classifying labeled data under different settings, visual evaluation of the full classified images does not support this conclusion. Figure 4.4 and Figure 4.5 show that MF tends to yield very blocky results, so samples on the class boundary are often misclassified. In Figure 4.4, the narrow river channel (blue pixels) is relatively small compared to the neighboring floodplain and the riparian pixels. The MF result shows that these water pixels are dominated by their local neighbors. In addition, some classes which are distributed according to complex texture-like spatial patterns are removed by the majority filtering process. A similar result over a different subset of the data is also shown in Figure 4.5. The individual class accuracy table (Table 4.3) provides additional insight. The table shows that MF performs very well on classes which are large in spatial extent, but fails to recognize small classes with complex boundaries. For example, MF performs well on woodlands and grasslands, two classes which are spectrally diverse due to variations in the density of vegetation, but are homogeneous in terms of the labels of their neighbors; however, the accuracy of MF is much lower for exposed soils due to the small sizes of the patches. These pixels were reclassified according to the label of their more frequently occurring neighbors. Images classified by MC-HSVM (Figure 4.4 and 4.5) do not have these problems. Overall accuracies of the extended edge test set achieved by MC-HSVM are consistently higher than the MF approach.

4.3.4 Comparing Classification Results from MRF-ICM and MC-HSVM

Algorithms based on MRF have been widely used for utilizing spatial-spectral information. Experiments performed during this study show that MC-HSVM is slightly better than MRF-ICM in terms of overall classification accuracies and is competi-

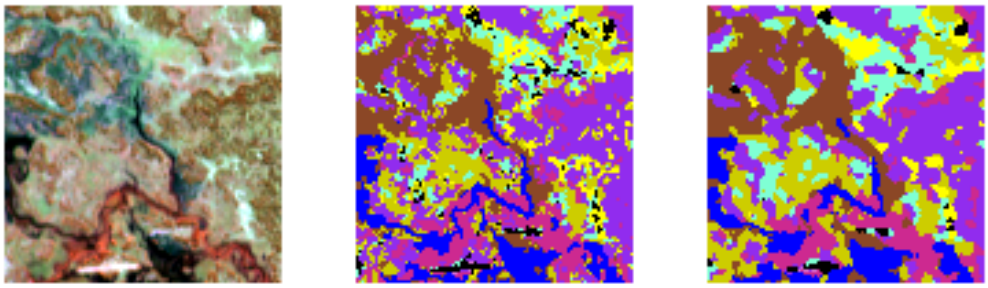

Figure 4.4: Wetland Area, (Left) Original Image, (Center) MC-HSVM Result, (Right) MF Result

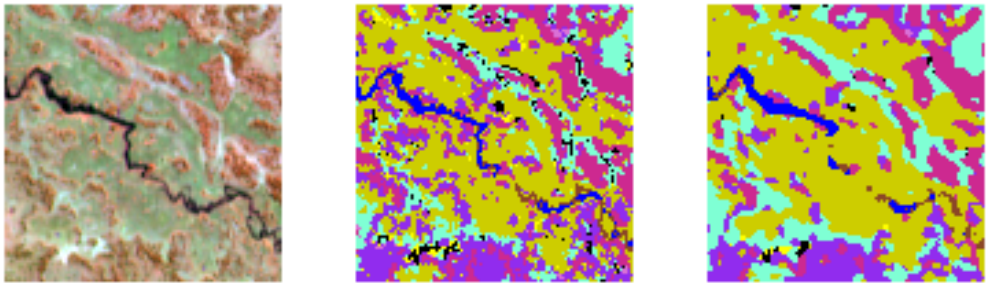

Figure 4.5: Wetland Area 2, (Left) Original Image, (Center) MC-HSVM Result, (Right) MF Result

tive with MRF-ICM in individual class accuracies. MC-HSVM has slightly higher accuracies for bushes and flat areas, like woodlands, grasslands and exposed soils, while ICM performs well in wetland classes, such as floodplains, firescars and island interiors. Because these wetland samples tend to have more complex geometric configurations, these results indicate that the capability of the MRF to represent the probabilities of neighboring classes is advantageous. The MC-HSVM method yields higher accuracies in homogeneous and small classes because it incorporates properly selected contextual information in the classification model.

Average processing times for the MRF-ICM and MC-HSVM classifiers are

quite different. The processing times for all four classifiers are listed in Figure 4.6. For the 40 Botswana experiments - each having 256*1465 pixels, 9 classes and 145 features using a 3GHz Pentium 4 CPU - the MRF-ICM is the slowest classifier, averaging 75 minutes per data set. MC-HSVM required 25 minutes to classify a data set, including computation for the max-cut stacking. This is only one-third of the processing time required by MRF-ICM. Thus, MC-HSVM is not only competitive with MRF-ICM in terms of accuracy, but is clearly superior to MRF-ICM with respect to computational effort.

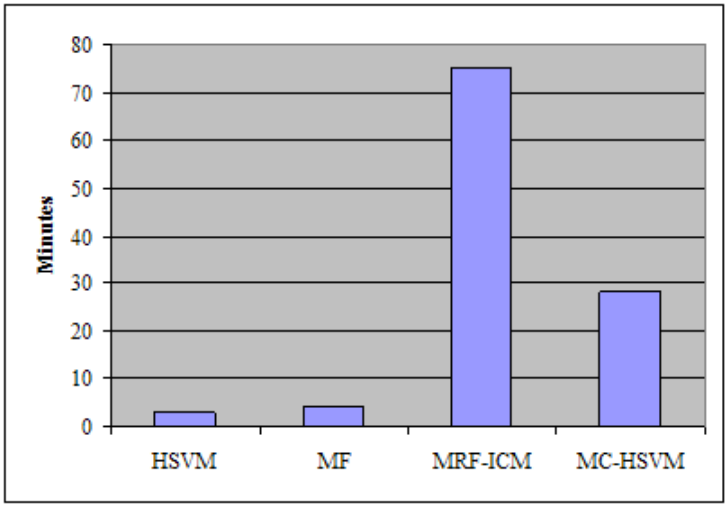

Figure 4.6: Average Processing Time for Classifying 40 Botswana Experiments

4.4 HSVM Based Ensemble Results

A mixture of expert algorithms incorporates randomization into the classification process, thus, increasing the domain covered by the classification process. It was observed in the previous study with random forests that the diversity of the ensemble is more critical than the number of classifiers in the ensemble for the overall classification accuracies [34] (also see Appendix B). Results presented in the previ-

ous section indicate that the majority filtering approach removes outliers from the homogeneous areas while the MC-HSVM approach helps recognize complex spatial boundaries. Since the two algorithms perform differently under different conditions, the performance of an ensemble that uses both classification algorithms was investigated in the Okavango Delta of Botswana because the region has a variety of contextual signatures.

4.4.1 Experiments with HSVM Ensembles

The max-cut set partition problem presented in Section 3.1.2 not only provides the boundary (cut) between homogeneous subsets but also the number of pixels, ranging from 1 to 9 for a second order neighborhood, in the subset in which the targeted pixel resides. This additional information provides an "index of homogeneity" for a spatial neighborhood. In this section, two types of ensembles are investigated - a simple ensemble and a switched ensemble. Experimental results are compared to those achieved by individual classifiers to investigate the benefit provided by the homogeneity index. The simple and switched ensembles are defined as follows:

1. Simple ensemble: Unlike the max-cut stacked vector approach which selects neighbors intelligently, the simple ensemble employs the five spatial patterns shown in Figure 4.7 to choose neighbors whose average spectral bands are used as stacked vectors. These five datasets are classified using HSVM. Class labels are assigned according to a simple voting process by the five classified maps.

2. Switched ensemble: The switched ensemble approach is based on the classification results of MF and MC-HSVM. In addition to the MF and MC-HSVM based algorithms, this ensemble uses the homogeneous index determined by the local set partitioning process. If there are any disagreements on the class label of a target pixel between MF and MC-HSVM, the homogenous index

decides which classifier gets the higher weight in the ensemble. A binary variable, I, whose value is determined by this homogeneous index is created. If the index is higher than a user defined threshold, then $I = 1$ otherwise $I = 0$. I indicates whether the classifier should agree with MF or MC-HSVM.

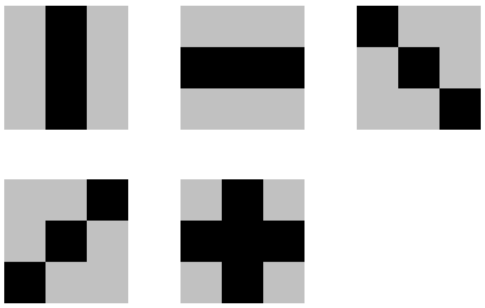

Figure 4.7: Five Spatial Patterns - Ensemble

These two ensemble approaches were applied to the same edge data set collected from the Botswana images described in Section 4.3. The overall classification accuracy is presented in Table 4.4, and the individual class accuracies are shown in Table 4.5.

Table 4.4: Botswana Edge Test Data: Ave. Accuracy (Std. Dev.) with Ensemble Results

Training %	MF	MC-HSVM	Ensemble	S-Ensemble
15%	84.5(2.41)	86.3(2.31)	86.2(2.25)	**86.7(2.20)**
30%	89.4(2.14)	89.9(1.49)	90.7(1.59)	**90.8(0.88)**
50%	89.8(1.25)	90.8(1.23)	91.2(1.20)	**91.9(0.87)**
75%	90.9(0.94)	91.9(0.66)	91.8(1.09)	**92.7(0.60)**

4.4.2 Discussion of Ensemble Results

In comparing the overall classification accuracies achieved by the simple ensemble method with those from the five different spatial patterns, there appears to be little

Table 4.5: Botswana Edge Test Data: Individual Class Ave. Accuracy (Std. Dev.) with Ensemble Results

Class	MF	MC-HSVM	Ensemble	S-Ensemble
Water	96.0(2.43)	98.6(0.60)	98.5(0.99)	**98.6(0.60)**
Primary Floodplain	94.3(3.56)	91.3(3.27)	**95.0(3.61)**	94.1(3.65)
Riparian	77.3(7.23)	80.0(3.05)	**81.3(6.20)**	77.9(2.37)
Firescar	**98.8(0.37)**	94.1(4.16)	96.9(1.67)	96.1(2.35)
Island Interior	**97.4(2.62)**	86.7(9.58)	95.3(2.38)	90.7(4.27)
Woodlands	98.5(0.35)	95.0(2.49)	**98.8(0.95)**	97.4(1.56)
Savanna	**99.1(0.58)**	97.6(1.17)	97.8(1.37)	98.4(0.67)
Short Mopane	**82.4(5.33)**	80.3(4.14)	80.1(6.23)	80.4(4.38)
Exposed Soils	66.1(5.59)	**93.2(3.62)**	77.6(4.31)	93.3(3.32)

gain in the average accuracies and a small reduction in the standard deviation of accuracies when the ensemble is used. Since spatial patterns of the Okavango Delta of Botswana are very complex, the simple ensemble has difficulty recognizing them. This weakness could possibly be easily corrected by using a slightly larger ensemble that has greater diversity but a longer processing time.

The new overall accuracy table shows that the switched ensemble takes advantage of both classifiers and gives the best classification results without the cost of obtaining a larger ensemble, particularly if it was based on randomization. This finding can be explained by a tie-breaker example. For the switched ensemble, whenever there are disagreements, the homogeneous index eventually becomes the tie-breaker. Because the index represents the homogeneity of a neighborhood, it provides a good indication of the classifier that performs best under such conditions. These results confirm the earlier conjecture. The individual class accuracies in Table 4.5 also support this finding. Results show that the switched ensemble method does better than MC-HSVM in the homogeneous areas and realizes greater improvements over MF in the mixed areas.

These additional experiments provide a better understanding about how to create an ensemble with two complementary classifiers. They also show that a homo-

geneous index is an excellent indicator for choosing the most appropriate classifier under certain circumstances, while avoiding the cost of building a large ensemble.

4.5 Summary

The goal of this part of the study is to fully exploit the spectral information provided by hyperspectral sensing, using contextual information to further improve classification and create a robust classifier. In this chapter, a knowledge-based stacking algorithm was developed using max-cut optimization to recognize the spatial class boundaries in the image. The proposed method was applied to data collected over the Okavango Delta of Botswana and compared to other very competitive and well studied approaches.

The classified map and classification accuracy tables presented in this chapter indicate that the proposed max-cut stacking method is able to provide more accurate predictions on both homogeneous areas and samples selected from mixed neighborhoods. It not only shows that it provides more detailed classification boundaries than the MF algorithm, but is also better than the MRF-ICM in both classification accuracy and speed.

Additional experiments using the switched ensemble demonstrate that the value of the homogeneous index obtained by the max-cut set partitioning process improves the overall classification accuracy. It provides a clear decision on which classifier should be assigned the higher weight, and as such, assigns the class label of novel samples. This new approach avoids the high cost of the linearly increasing processing time of a larger ensemble with a simple indicator obtained from currently designed procedures. This intuitive classifier also requires less human input, satisfying the second overall goal of this research.

Chapter 5

Learning the Shortest Path Network for Knowledge Transfer

Classification algorithms presented in the previous two chapters are shown to perform efficiently under the condition that training and testing samples follow the same set of class distributions. Unfortunately, such an assumption is not always valid when image data are acquired at multiple times, or testing samples are collected from spatially disjoint regions as in the knowledge transfer scenario. An algorithm that adapts changes of class distributions over extended regions or time is required to overcome the knowledge transfer problem. Unfortunately, as the decision boundaries used to classify the hyperspectral images become more complex, the generalization error eventually increases because of over-training [68]. Although ensemble methods presented in Section B.3 alleviate this problem by reducing the model variance, they are computationally costly due to the large number of classifiers (50-100) required in the ensemble [34]. Strong classifiers such as SVM also do not typically perform well in a knowledge transfer setting. It is important to

develop a simple classifier that can adapt to such changes, as well as to maintain good classification accuracies for the training and testing data.

5.1 Background to Nonlinear Learning

Nonlinear manifold learning algorithms transform data to a new space based on the pairwise distances between samples using local search methods. Methods such as Isometric feature mapping (Isomap) [75] and local linear embedding (LLE) [67] assume that the original high dimensional data actually lie on a low dimensional manifold defined by local geometric differences between samples. Isomap was recently applied to hyperspectral data by Bachmann *et al.* [3, 4]. Their results indicated that the transformed data of Isomap produced more meaningful features than those obtained by the maximum noise fraction (MNF) transform [33] in visual analysis. However, computational requirements precluded direct application to large remotely sensed data sets. Although manifold learning algorithms focus on nonlinear dimension reduction and representation of high dimensional observations, they also provide opportunities for classification of hyperspectral data because they better represent the nonlinear phenomena in the data and provide increased class separation. It is conjectured that better representation of the physical phenomena may also lead to improved knowledge transfer in classification.

The first half of this chapter focuses on the effect of the distance updating scheme and its impact on solving the knowledge transfer problem. A discussion of Isomap, that includes the shortest path algorithm and multidimensional scaling (MDS), is contained in Section 5.2.1. The new proposed shortest path k-nearest neighbor classifier (SkNN), which is closely related to the shortest path updating scheme between labeled and unlabeled samples, is also described. Results of dimension reduction for a test site in Botswana and comparisons of classification accuracies achieved by SkNN and other competitive classifiers are presented in Section

5.3. With these initial results, the second half of this chapter starting from Section 5.4 involves the investigation of approaches to increase the speed of SkNN when the number of samples is large.

5.2 Isomap and Shortest Path k-nearest Neighbor

This section contains an overview of Isomap and describes how shortest path networks adapt changes of spectral signatures between different data sets. Details of the proposed SkNN classifier are also presented.

5.2.1 Isometric Feature Mapping (Isomap)

Isomap nonlinear manifold learning is based on shortest path network updating and multidimensional scaling (MDS). The original input, $\mathbf{X} \in \Re^{d \times n}$, representing n samples and d dimensions, is first used to calculate the pairwise distances within a user-defined neighborhood. A shortest path algorithm is applied to update those pairwise distances beyond the neighborhood. The updated distance matrix is used by MDS to evaluate the true dimension of the manifold, which should lead to higher classification accuracies.

Shortest Path Network

Isomap employs a user-defined neighborhood and the shortest path algorithm to discover the manifold. It first defines K_i, the set of neighborhood nodes of node i, to create a distance matrix $\mathbf{D}'$ with elements d_{ij}. If $j \in K_i$, $d'_{ij} = d_{ij}$. If $j \notin K_i$, $d'_{ij} = \infty$. Isomap then accumulates the distance beyond the set K_i along the shortest path to obtain $\mathbf{D}_{stp}$. This idea can be explained using the following Swiss roll example. Figure 5.1 shows that the Swiss roll resides in a three-dimensional space. By using a properly defined neighborhood, these samples can be unrolled and represented in a flat two-dimensional image in Figure 5.2.

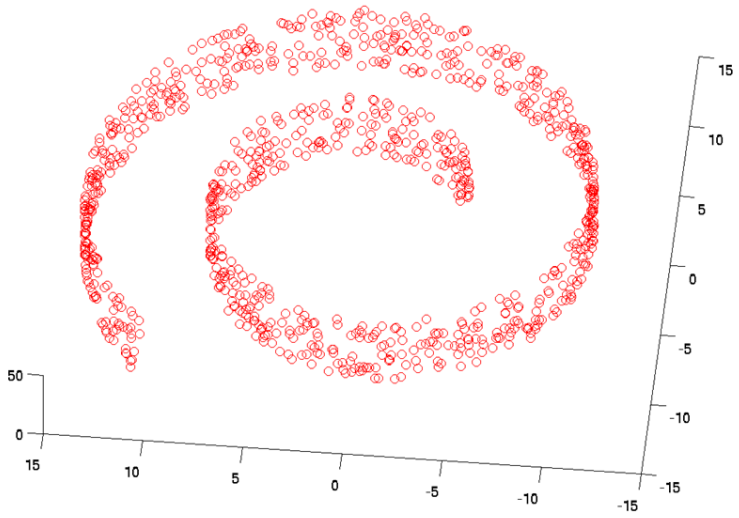

Figure 5.1: Swiss Roll Example: Original 3-Dimensional Data

The shortest path network is constructed from a directed graph $G = (N, E)$, where N represents the set of n nodes, and E represents the edges of the graph. The value of d_{ij} represents the length (cost) of E_{ij}, while x_{ij} is the amount of flow from N_i to N_j. The shortest path algorithm finds the paths from a root node N_1 to all other nodes to minimize the sum of the individual path lengths. This problem is formulated as a network flow programming problem:

$$\begin{aligned} \min \quad & z = \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} x_{ij} \\ s.t. \quad & \sum_{j=2}^{n} x_{1j} = n - 1 && (5.1) \\ & \sum_{j=1}^{n} x_{ij} - \sum_{j=1}^{n} x_{ji} = -1, \quad i = 2, ..., n && (5.2) \\ & x_{ij} \geq 0, i \neq j = 1, ..., n. && (5.3) \end{aligned}$$

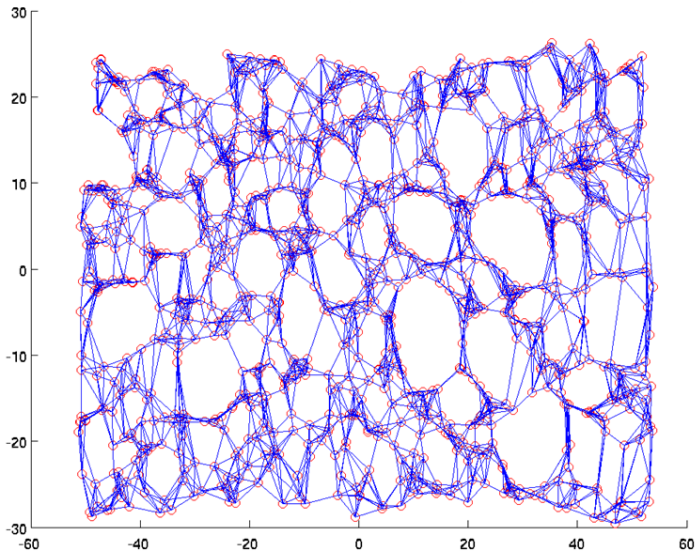

Figure 5.2: Swiss Roll Example: After Transformation via Isomap

In this optimization problem, Equation (5.1) is the supply at the root node, (5.2) represents conservation of flow, and (5.3) is the non-negativity constraint. Because this is a pure network flow problem, it can be modeled as a linear programming problem and solved either via the simplex method, guaranteeing an optimal integer solution [78]. Isomap solves the problem via a simple, computationally efficient algorithm developed by Dijkstra [25] [1].

Multidimensional Scaling

Multidimensional scaling (MDS) is a linear dimension reduction technique that places a set of samples in a meaningful dimensional space that explains the similarity between samples. Given a distance matrix $\mathbf{D}$, and assuming that a lower dimensional input $\mathbf{Y} \in \Re^{l \times n}, l \ll d$ exists such that $\delta_{ij}^2 = ||\mathbf{y}_i - \mathbf{y}_j||^2 \approx d_{ij}^2$ ($\mathbf{y} \in \Re^{l \times 1}$ represents a sample with a lower dimension) and $\mathbf{Y}_i$ are orthogonal, it can be shown

[1] For more details, see http://www.cs.utexas.edu/users/EWD/

that $\mathbf{Y}$, calculated by classical MDS, is equivalent to a vector of the first l principal components of $\mathbf{X}$ if the Euclidean pairwise distance matrix is used [70]. Here, MDS is used to evaluate the true dimension of $\mathbf{D}_{stp}$.

Experiments in [75] demonstrated that $\mathbf{D}_{stp}$ is able to define the nonlinear manifold, and that it can be represented globally by MDS in a lower dimensional space. For example, if the pairwise distances between a set of 100 cities of the US are represented by MDS, a three dimensional space is required to preserve the pairwise relationships between these cities globally. If the distance is updated locally and nonlinearly so that only distances between cities of a defined neighborhood are considered, these cities would lie on a two dimensional map.

5.2.2 Shortest Path k-Nearest Neighbor Classifier

If high dimensional data can be preserved in a low dimensional manifold, an updated distance matrix, which preserves the local information on a graph while increasing the distances between non-neighbor samples, should be useful for classification. In addition, if no labeled samples are available from the new areas, the shortest path updating approach provides a framework for the original model to gradually adapt to spectral changes between the new and original areas.

In this study, $\mathbf{D}_{stp}$ is investigated for land cover classification. Figure 5.3 and Figure 5.4 show that shortest path updating moved similar samples closer to each other, while dissimilar points were more separated, indicating the potential usefulness of the Isomap approach. ‘

If a set of samples can be presented in a low dimensional space, the simple k-nearest neighbor classifier is often the most competitive algorithm for classification. Given a novel observation, kNN classifies it according to the class label of its k nearest neighbors, in the distance sense. The kNN has several advantages. The method is easy to implement, its classification accuracy is very good on low dimen-

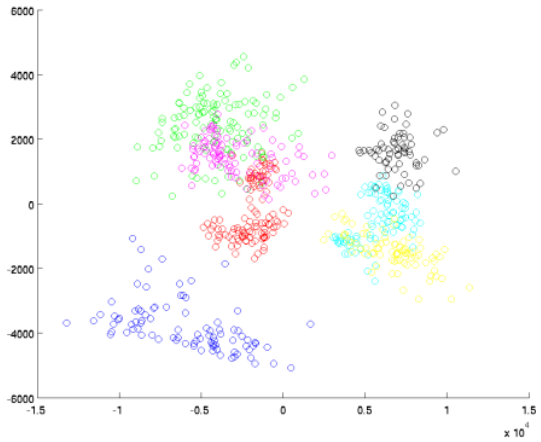

Figure 5.3: Two dimensional PCA plot, 8 Classes (exclude water), Hyperion Data of Botswana

sional problems, and it provides nonlinear decision boundaries. Furthermore, kNN handles multi-class problems, the norm in land cover classification problems.

A new method, the shortest path k-nearest neighbor classifier (SkNN), is proposed to utilize the information learned from the low dimensional nonlinear manifold. SkNN approximates each spectral signature as a probability distribution and uses

$$d_{ij} = \frac{1}{2} \sum_{\forall x} \left(f_i(x) \log \frac{f_i(x)}{f_j(x)} + f_j(x) \log \frac{f_j(x)}{f_i(x)} \right) \tag{5.4}$$

the average Kullback-Leibler divergence [45] between the spectral signatures of sample i and sample j as the distance measure. This KL-distance matrix $\mathbf{D}$ is converted to $\mathbf{D}_{stp}$, as described in Section 5.2.1. The k-nearest neighbor algorithm then classifies the unlabeled samples projected in the space of the new distance matrix $\mathbf{D}_{stp}$.

5.3 SkNN Results

The benefits of applying shortest path networks to hyperspectral data are evaluated in terms of dimension reduction and classification accuracy of the Hyperion data

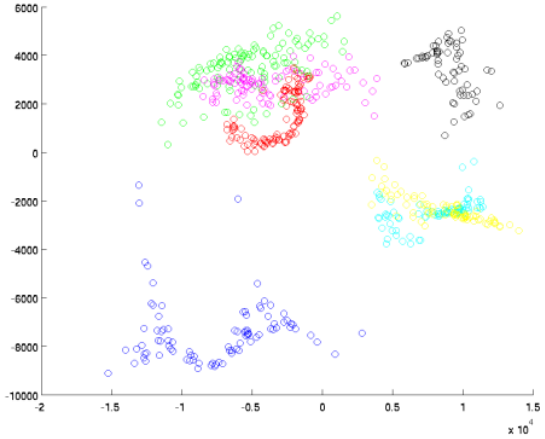

Figure 5.4: Two dimensional Isomap plot, 8 Classes (exclude water), Hyperion Data of Botswana

acquired over the Okavango Delta, Botswana in 2001. Details of this data set are described in Appendix A.2.2.

5.3.1 Dimension Reduction

Although studies have shown that methods such as principal component analysis (PCA) can reasonably discover the true structure of the 150+ bands in hyperspectral data on a linear subspace, the original high dimensional data may lie on a nonlinear manifold. To evaluate the true dimension of the manifold, SStress [74]:

$$ss = \left[\frac{\sum \sum_{i<j} (d_{ij}^2 - \delta_{ij}^2)^2}{\sum \sum_{i<j} d_{ij}^4} \right]^{\frac{1}{2}} \tag{5.5}$$

is used by MDS to evaluate the similarity of $\mathbf{D}_{stp}$ and Δ_l, where d_{ij} is an element of D, Δ is the reconstructed distance matrix, and l is the dimension of $\mathbf{Y}$. The value of SStress is always between 0 and 1. Any value less than 0.1 is considered to indicate good representation in the given l dimensions. Values in Table 5.1, indicate that SStress became less than 0.1 when $l \geq 2$. Thus, Isomap found the embedding

manifold of this Hyperion data could be represented in a very low dimensional space with little loss of information.

Table 5.1: Botswana Data: SStress with l dimensions

l	1	2	3	4
SStress	0.167	0.046	0.042	0.046

5.3.2 Classification Results: Test Data and Spatially Disjoint Areas

This study uses the same sampling approach described in Section 3.2.2 and has ten samples for each of four sampling rates (75%, 50%, 30% and 15%). Because the training and test data are spatially collocated, the spatially disjoint (SD) test set described in Section 3.2.3 was also used to evaluate the generalization of these classifiers to another area. Note that this extended data may have substantially different characteristics as it was collected from a geographically separate location. The goal here was to investigate each method's capability of extending the results obtained from one area to another where data are not so spatially correlated with the original training samples. More details of this dataset are available in Appendix A.2.2.

For comparison, experiments were performed using the best basis binary hierarchical classifier (BB-BHC) with weighted prior [46], hierarchical SVM (HSVM) [14], k-nearest neighbor (kNN) on the original space, and the proposed shortest path k-nearest neighbor (SkNN). The average test data classification accuracies and their corresponding standard deviations for the 10 experiments conducted with each classifier are listed in Table 5.2. Here, k = 5 was chosen by a cross-validation scheme. In cross-validation experiments, the SD of the test accuracies increased slightly, but test accuracies decreased when k was increased. Results were also obtained by kNN using $3 \sim 5$ PC bands, but the resulting accuracies were consistently lower than those achieved by the kNN classifier applied to the original space.

Table 5.2: Botswana Test Data: Ave. Accuracy (Std. Dev.)

Training %	BB-BHC	HSVM	kNN	SkNN
15%	92.6(2.16)	**96.5(0.95)**	83.2(3.17)	94.2(1.19)
30%	95.5(1.68)	**97.3(1.14)**	91.3(1.74)	96.4(1.33)
50%	97.6(0.74)	**97.9(0.51)**	95.0(1.27)	97.1(1.24)
75%	**98.1(0.60)**	97.7(0.51)	96.1(1.24)	97.5(0.81)

The overall trend shows that classification accuracies of the test set for this data set increase as the size of the training sample increases for all four classifiers. HSVM achieves the highest overall average accuracies, although all classifiers perform well at 15% sampling rate, indicating that they are robust to small samples of test data.

Classification accuracies on the spatially disjoint (SD) test set are contained in Table 5.3. HSVM performed consistently well on the test set, but not on the

Table 5.3: Botswana Spatially Disjoint (SD) Test Data: Ave. Accuracy (Std. Dev.)

Training %	BB-BHC	HSVM	kNN	SkNN
15%	84.1(2.70)	80.7(3.1)	77.3(2.22)	**84.7(2.14)**
30%	84.3(1.13)	84.1(1.57)	79.8(1.24)	**86.1(2.60)**
50%	84.5(1.24)	84.1(0.83)	81.3(0.86)	**86.8(2.06)**
75%	85.8(0.60)	84.6(0.61)	82.2(0.60)	**87.5(1.06)**

SD test set. Because the spectral characteristics of the training and test data were different from those of the SD test set, this supported the notion that while SVM is a strong classifier, it might not be robust with respect to changes in data characteristics. Although BB-BHC was competitive with SkNN, as indicated by the relatively low variance of accuracies obtained on the SD test set, the average accuracy of the individual classes ranges from 70-100% for the BB-BHC, and 80-100% for the SkNN, while ranges of the standard deviations of the respective accuracies are 1.8-5.6 and 2.8-10, respectively. SkNN achieves higher average accuracies but

with larger standard deviations because it is more sensitive to samples. Since $\mathbf{D}_{stp}$ evolves as new samples are included in the distance matrix, SkNN not only performed well on the test set but also produced the highest accuracies on the SD test set at all four sampling rates. Results achieved by kNN are included to demonstrate that SkNN realizes benefits from the implementation of shortest path updating. For the Botswana experiment that had 790 samples, 9 classes and 145 feature spaces, using a 3GHz Pentium 4 CPU machine, kNN finished training and testing in 31 seconds, while HSVM required 40 seconds. BB-BHC required 65 seconds and the proposed SkNN required 149 seconds of CPU time.

5.3.3 Classification: Multi-temporal Data

Solving a knowledge transfer problem in a temporal framework involves using training models developed from data acquired at an earlier time to classify images obtained at a later time. Thus, two additional Hyperion images acquired over the Okavango Delta, Botswana on June 16th and July 1st of 2001 were processed, and additional labeled samples were selected for multi-temporal experiments. During this period, vegetation changed gradually, and the flood on the Okavango advanced. The original May 31st data set and two newly processed data sets are referred to as May, June and July data sets respectively. Three additional experiments, May-to-June, May-to-July and June-to-July, were designed for this multi-temporal setting to evaluate the impact of classification of data sets acquired over different time intervals in a dynamic environment. Classification accuracies obtained by the proposed SkNN classifier, BB-BHC, and HSVM are presented in Table 5.4, 5.5 and 5.6 respectively.

Accuracies achieved in this multi-temporal classification scenario did not improve as percentages of available labeled samples increased. This trend is different from that shown in Table 5.2, for the traditional test data seems to indicate that

Table 5.4: Botswana Test Data (May to June): Ave. Accuracy (Std. Dev.)

Training %	BB-BHC	HSVM	SkNN
15%	57.6(2.16)	66.2(0.95)	**70.2(1.19)**
30%	53.2(1.68)	60.2(1.14)	**71.8(1.33)**
50%	55.6(0.74)	62.4(0.51)	**70.1(1.24)**
75%	56.0(0.60)	65.2(0.51)	**68.6(0.81)**

Table 5.5: Botswana Test Data (May to July): Ave. Accuracy (Std. Dev.)

Training %	BB-BHC	HSVM	SkNN
15%	58.9(2.70)	64.4(3.10)	**71.9(2.14)**
30%	61.9(1.13)	66.4(1.57)	**73.7(2.60)**
50%	65.2(1.24)	68.8(0.83)	**70.6(2.06)**
75%	68.3(0.60)	68.8(0.61)	**69.1(1.06)**

increasing the number of labeled samples results in stronger classification boundaries and a resulting loss of flexibility for each classifier. The comparison of BB-BHC and HSVM shows that best bases feature extraction forces BHC to fit the original class distribution and does not provide the same level of robustness as that provided by HSVM. Although accuracies achieved by SkNN have slightly higher standard deviations, their classification accuracies are consistently the highest among these three classifiers for this data set. These experiments indicate that the shortest path networks adapt to changes in spectral signatures of data in a multi-temporal setting and provide promising results when applied to hyperspectral data.

With these initial results, the second half of this chapter involves investigation of approaches to reduce the computational demands of SkNN.

5.4 Applying Nonlinear Manifold to Extended Areas

Although the results from the previous section indicate that Isomap can represent nonlinear information and characterize hyperspectral data in low dimensional

Table 5.6: Botswana Test Data (June to July): Ave. Accuracy (Std. Dev.)

Training %	BB-BHC	HSVM	SkNN
15%	84.1(0.85)	86.9(1.20)	**88.6(1.10)**
30%	81.6(0.83)	85.9(0.84)	**89.8(1.10)**
50%	82.6(0.58)	87.8(0.57)	**90.2(0.67)**
75%	83.5(0.76)	85.2(0.69)	**90.4(1.15)**

spaces, the shortest path updating scheme creates a bottleneck in the local search process as the complexity of the inter-pixel network increases exponentially with image size. Remote sensing images typically contain a very large number of pixels, so divide-and-conquer approaches were utilized in these two studies [3, 4, 15]. A tiling method that adjusted the eigenvalue/vector pairs of each image tile to form a unique map reduced computation time when Isomap was applied to large-scale images [3]. In my previous research [15], building the global Isomap was avoided by allowing disconnected subgroups in the manifold and using the k-nearest neighbor (kNN) method as the base classifier. Because kNN classifies samples only according to the updated geometric distance between labeled and novel samples, a unified re-projected map is not required. Use of a "localized" map dramatically reduces the number of connected edges and prevents searches of certain nonessential shortest paths. Although the overall mean classification accuracy was high for large areas, kNN is sensitive to outliers, which resulted in high standard deviations of the classification accuracies. The section that follows describes a new landmark point selection approach that reduced computation in the shortest path updating scheme and creates a well connected map so that a more robust classifier can be utilized to reduce the standard deviation of classification accuracies.

5.4.1 Landmark-Isomap and Landmark Points Selection

Experiments in Section 5.3 demonstrate that $\mathbf{D}_{stp}$ is able to define the nonlinear manifold, and that it can be represented globally by MDS in a lower dimensional space. However, although Dijkstra's algorithm is efficient for finding the shortest path from a root node to the rest of the nodes, building the whole shortest path network, with an order of $O(k|N^{|}2\log|N|)$, is problematic when the total number of samples, $|N|$, is large. In order to develop a more robust classifier that still exploits the advantage of nonlinear dimension reduction and is computationally competitive, the landmark Isomap (L-Isomap) [22] was investigated. L-Isomap is identical to Isomap except that it uses a subset of points to build the map. To eliminate unnecessary calculations and speed up the shortest path search process, L-Isomap randomly selects n landmark points from the original data to construct its manifold [22]. Instead of building an $N \times N$ shortest path network, L-Isomap uses a much smaller $n \times N$ network, which requires less effort. MDS operations are also reduced on this network. Samples that are not selected for landmarks are placed on the manifold via the derived embedding vectors and their updated distances to n landmark points. Experiments indicate that if samples are equally distributed on a smooth manifold, the L-Isomap is capable of achieving the same level of data compression as the Isomap without losing too much information.

Because pixel spectral signatures of different land cover types are often not located in spectrally contiguous clusters, the original L-Isomap fails to reconstruct the low dimensional manifold when it is applied to hyperspectral data for dimension reduction. Results of preliminary experiments performed in this study show that if samples are assumed to lie on a manifold that has k facets and can be represented by k clusters, extreme points perform better as landmark points in preserving the manifold than random points or cluster centers (facets). For example, in order to reconstruct a human's face (a manifold), points that are close to the boundaries of

each facet, such as the tip of the nose, chin and the dip between two eye brows should be selected as landmark points instead of cluster centers. There are many ways to find such extreme points. In this study, a minimum spanning tree cut (MST-cut) is used to locate these landmark points.

5.4.2 Minimum Spanning Tree

Assume a connected, undirected graph $G = (N, E)$. A spanning tree is a graph that connects all N nodes and has no cycles. The minimum spanning tree is a tree with the lowest total cost. There are two advantages in the use of MST for finding landmark points for nonlinear embedding:

1. Existing algorithms can solve MST in polynomial time and can provide the optimal solution.

2. MST is unique for a given graph G, unlike the shortest path tree that is unique for each root node.

Since MST is closely related to single linkage hierarchical clustering [32], partitioning by cutting the heaviest edge in the MST provides two boundary samples that come from two clusters (facets). These boundary samples are chosen as landmark points to reconstruct the manifold.

In addition to using landmark points on the boundaries to create the manifold, the proposed approach calculates the distances from a novel sample to its k-nearest landmark points. It then uses that distance information to locate the novel sample's position on the manifold. Such an embedding approach is widely used in global positioning systems (GPS). It not only speeds up the re-projecting process, but is also able to accommodate a manifold that has many facets and samples that are disjoint. Results of SStress and classification accuracy from multiple experiments are presented in the next section to support this finding.

5.5 L-Isomap Results

The benefits of applying the new L-Isomap to hyperspectral data were evaluated relative to manifold reconstruction and classification of the Hyperion data that were presented in the earlier section.

5.5.1 Manifold Reconstruction

The different landmark point selection methods were compared using their updated distance matrices. SStress (see Equation (5.5)) was employed to evaluate the similarity of $\mathbf{D}_{stp}$ and $\mathbf{D}_l$, where l includes random selection, k-means clustering centers and the proposed MST-cut approach for selecting landmarks. Results are obtained for 4 sampling rates (50%, 25%, 12.5% and 10%), each with 10 runs, to compute the estimated standard deviation. The value of SStress is always between 0 and 1. Any value less than 0.2 is considered to indicate good representation. Table 5.7 shows

Table 5.7: Botswana Data: Ave. SStress (Std. Dev.)

Sampling Rates	Random	K-means	MST
50%	0.42(0.05)	0.39(0.04)	**0.20(0)**
25%	0.49(0.03)	0.47(0.03)	**0.21(0)**
12.5%	0.55(0.01)	0.54(0.01)	**0.30(0)**
10%	0.57(0.01)	0.55(0.01)	**0.32(0)**

that the MST algorithm is consistently the best among these three methods at all four different sampling rates. Because landmark points collected by MST-cut are the same for all ten experiments, it has a zero standard deviation. The following supports our contention: points on the edges are better representatives than cluster centers or randomly selected points when used to reconstruct a manifold created by the whole dataset.

5.5.2 L-Isomap Classification

The proposed algorithm was applied to the same Botswana data set presented in Section 5.3 and also in Appendix A.2.2. The 50% sampling rate data are used in the experiments, and methods are evaluated using the ten test samples composed of 25% of the original training data. Because the training and test data are spatially collocated, the spatially disjoint test set collected from a geographically separate location and described in Appendix A.2.2 was used to evaluate the generalization of these classifiers to another area. Data were processed by Isomap and L-Isomap (which uses 25% of the total samples) to perform feature extraction. The original 145 bands were re-projected to the first five MDS bands. These data sets were trained and classified by a set of classifiers to investigate how landmark points impact the overall classification accuracy.

Experiments were performed using the k-nearest neighbor, C4.5 decision tree [61], logistic regression [28] and linear kernel SVM on the low-dimensional input space determined by Isomap and L-Isomap. The average test data classification accuracies and their corresponding standard deviations for the 10 experiments conducted with each classifier are listed in Table 5.8. The overall trend shows that

Table 5.8: Botswana Test Data: Ave. Accuracy (Std. Dev.)

Training: 50%	kNN	C4.5	Logistic Reg.	SVM
Isomap	91.4(2.51)	90.7(1.79)	91.6(1.98)	**92.2(1.19)**
L-Isomap	88.6(1.54)	87.4(1.48)	87.8(2.59)	**89.1(1.31)**

stronger classifiers such as SVM and Logistic regression achieve somewhat higher classification accuracies than kNN and the C4.5 decision tree. The low-dimensional input space defined by Isomap gives consistently better results than those obtained by L-Isomap. Because the spatial signatures of training and testing samples have the same distribution, the new input space defined by landmark points (L-Isomap)

can never achieve higher classification accuracy than Isomap.

Classification accuracies on the spatially disjoint (SD) test set are contained in Table 5.9. Because the spectral characteristics of the train/test data are differ-

Table 5.9: Botswana Spatially Disjoint (SD) Test Data: Ave. Accuracy (Std. Dev.)

Training: 50%	kNN	C4.5	Logistic Reg.	SVM
Isomap	77.5(2.13)	76.9(3.32)	79.0(2.39)	**80.7(2.89)**
L-Isomap	76.9(1.79)	76.4(1.67)	**79.5(2.75)**	78.2(1.79)

ent from those of the SD test set, the embedding approach, which only considers distances from k-nearest landmark points, helps narrow the accuracy gap between Isomap and the proposed MST-cut L-Isomap. The proposed L-Isomap even provides higher accuracy than Isomap achieves when trained and classified by a logistic regression classifier. Similar to results from the test set, stronger classifiers do well in these experiments, contradicting the previous finding in [14] that reveals SVM to be better able to adapt to signature changes in this data set. The reason for this is that the nonlinear manifold reshaped the input space so that it became more adaptable to changes. It handled the updating and provided a better exploitation of the powerful SVM classifier. Since the processing time is reduced, it becomes plausible to apply Isomap as a pre-processor for dimension reduction and to apply classifiers other than the k-NN to the new data set. For the Botswana experiment that had 790 samples, 9 classes and 145 feature spaces, the Isomap feature extraction, using a 3GHz Pentium 4 CPU machine, required 139 seconds of CPU time while L-Isomap reduced that to 40 seconds. Time spent on training and testing was the same for both inputs with kNN requiring 9 seconds, C4.5 needing 19 seconds, logistic regression taking 45 seconds and linear-SVM requiring 31 seconds.

5.6 Summary

This study investigated of the advantages and weaknesses of the Isomap and L-Isomap when applied to hyperspectral data. Evaluations of dimension reduction and representation of high dimensional observations by Isomap and L-Isomap were conducted. This study also included an investigation of L-Isomap in conjunction with classification of hyperspectral data. The proposed MST-cut landmark selection approach was compared to random selection and k-means cluster centers.

The results show that the shortest path network learns and preserves the pairwise distances of labeled samples and gradually updates the network when unlabeled samples are added into the model. This approach is useful for the knowledge transfer problem without utilizing any labeled data from the new regions. In addition, these results indicate that the samples on cluster boundaries are better landmark point candidates that also preserve the manifold created by the whole sample set. Thus, these landmark points can be used initially to construct a similar manifold with significantly less processing time. The new embedding process makes applying non-linear manifold learning on large data sets possible. Stronger classifiers such as SVM can be applied to the modified data set to achieve both high classification accuracy and robustness.

Chapter 6

Conclusions and Future Work

The launch of Hyperion, the first space-based hyperspectral satellite, ushered in the era of globally available, inexpensive hyperspectral data. Although promising applications are numerous, particularly those involving classification, expert knowledge is necessary for the extensive manual tuning required by most methods to fully exploit the data's potential. Unfortunately, researchers who can make the best use of these data are often not experts in data analysis methodology. Additionally, tracking the continuously changing earth environment necessitates multi-temporal studies which in turn require n times the original cost to handle n scenes of multi-temporal data. This study focused on developing advanced tools for land cover classification which are computationally advantageous and largely automated. By providing a lower entrance knowledge threshold, this research will hopefully facilitate utilization of hyperspectral sensors. Motivated by these challenges, this study proposes a series of knowledge-based learning processes to make land cover classification more efficient and ascertainable for future researchers.

6.1 Summary of Contributions

This research, whose primary application is classification of hyperspectral data for land cover mapping, focused on developing machine learning methods for analyzing high dimensional problems that increases classification accuracies while reducing computation and the need for human-based tuning.

This research focused on three primary research areas. The resulting contributions are listed in the following sections.

6.1.1 Limited Number of Labeled Samples

Since the launch of the EO-1 satellite, researchers in the remote sensing community have made tremendous strides in developing advanced methods that facilitate the utilization of data from hyperspectral sensors. The majority of these efforts focus on developing algorithms that handle the hyperspectral sensors' high dimensional feature space or semi-supervised classification algorithms that effectively increase the number of labeled samples and avoid the "curse of dimensionality" for parametric classifiers such as Maximum Likelihood (ML) or Fisher's linear discriminant. In addition, nonparametric classifiers became attractive because of their relative insensitivity to high dimensional inputs. Among the many nonparametric classifiers, support vector machines (SVM) were shown to be the particularly promising for achieving high classification accuracies. However, while SVM classifiers are very powerful classification algorithms, they also pose challenges. For example, the SVM framework is binary and requires an output space decomposition algorithm to handle multi-class problems.

Having a framework for multi-class problems that integrates with SVM is important; thus, the classification process must not only provide high accuracies but also must require a limited amount of tuning. This research addressed this problem through the development of integrated hierarchical support vector machines

that help reduce the complexity of the decision boundaries at each internal node of the hierarchy. The proposed max-cut decomposition framework was coupled with the SVM to create a powerful classifier that maintains natural class affinity. Results presented in Section 3.2 demonstrate that the integrated hierarchical support vector machine (HSVM) performs well on hyperspectral data when limited training samples are available and dramatically reduces the time required by the tuning process.

6.1.2 Utilizing Contextual Information

Because of difficulties in obtaining adequate labeled data to represent a sufficient amount of contextual information, popular Markov random field (MRF) approaches suffer the "curse of dimensionality." As a result, hyperspectral classifiers do not typically exploit spatial information. Since the location of a pixel within a homogeneous region or on a boundary has important consequences for classification, a max-cut stacking method that exploits both spatial and spectral information is introduced in this study. The new process improves classification of pixels whose spectral values reside in the tail of the class distribution, pixels near class boundaries, or possibly of mixed pixels. The classified images and classification accuracy tables shown in Section 4.3 support the value of the new methodology.

In addition, this study demonstrated that max-cut stacking performed the best of the spatial-spectral classifiers in mixed areas, while majority filtering (MF) produced better results in homogeneous areas. A weighted ensemble approach was proposed to take advantage of this finding. Additional experiments showed that combining the advantages of both through the weighted ensemble of the max-cut HSVM (MC-HSVM) and MF increased classification accuracies of complex landscapes by removing outliers of homogeneous areas and identifying boundary pixels.

6.1.3 Solving Knowledge Transfer Problem

The third contribution of the study related to knowledge transfer for classification of areas where is the quantity of training data is limited, or none is available.. While this problem was considered in both HSVM and the MC-HSVM frameworks, the most successful developments involved manifold learning. For knowledge transfer, the shortest path network is the most critical aspect as it is used to update the pairwise distances between both labeled and unlabeled samples. Since the shortest path algorithm (STP) preserves the distances (dissimilarities) between similar samples and extends distances between samples that are spectrally different, the updated distance matrix transfers the knowledge obtained from one region to the new region with no labeled samples. The classification accuracies achieved by the proposed shortest path k-nearest neighbor (SkNN) classifier for multi-temporal experiments demonstrate the value of the approach.

An issue of the lengthy CPU time required to develop the shortest path network was also addressed. The majority of pairwise distances between millions of pixels in an image are irrelevant to the final classification model. To reduce the time required for the STP process, an intelligent landmark points selection algorithm that collects samples on the spectral boundaries between classes was proposed in Section 5.4.1. The process significantly reduces the time required to create a shortest path network and preserves more information than randomly selected landmark points and landmark points defined by clustering centers. The intelligent landmark points algorithm deserves further study with respect to creating disjointed STP networks.

6.2 Future Work

Algorithms presented in this study serve as a platform for future research. Since many current studies are devoted to the limited labeled sample problem, future

researchers should focus on the utilization of spatial information and the adaptation of knowledge from one region to a new region.

6.2.1 Pyramid Landmark Point Selection

To speed up the shortest path updating approach presented in Section 5.2.1, a multi-resolution pyramid approach could be helpful for constructing the manifold of a hyperspectral image. Local manifolds would be defined over subsets of the image at its lower levels. Data would be subsampled, and manifolds incrementally developed over more extended regions at higher levels of the pyramid. The pyramid approach could provide two advantages. First, a relatively small number of samples is required to build a manifold at the top of the pyramid. These samples could serve as landmark points at the lower level blocks of the global pyramid. Shortest sample paths inside the block could be calculated with respect to the landmark points so that the computation complexity would be reduced. Secondly, the pyramid approach should be helpful in introducing localized spatial and contextual information. Manifolds would reside over a given block of the pyramid at a given level and correspond to a spatial location within the image, so the spatial/contextual information would be transferred across the levels.

6.2.2 Utilizing Contextual Information

Selection of labeled samples is one of the most time consuming processes involved in land cover classification. It requires a skilled user who can recognize ground truth through the utilization of a combination of high resolution images, color maps and the understanding of a region. These abilities come from years of experience and are difficult to transfer to new or junior researchers. In addition, collecting these samples pixel by pixel is a long and tedious process. Since max-cut binary segmentation separates a spatial neighborhood into homogeneous subsets, it could simplify

the collection process by creating smaller, homogeneous patches. Researchers could assign labels to these patches instead of individual pixels. This procedure should reduce human errors and the processing time of the labeled sampled selection process. The unsupervised image segmentation process would also benefit researchers who are new to a study area by providing a quick summary of the region and by providing a training tool. Eventually, having these homogeneous image segments would benefit land cover classification through the simplification of the selection of labeled samples.

Contextual information can also be used for developing a post-processing approach which accounts for neighborhood configurations (similar to MRF) to correct assignments of classes that are spectrally similar but not geographically collocated. For example, island interiors have very similar signatures to exposed soils but have quite different neighbors. Island interior regions are often located in the wetland area, while exposed soils are often mixed with grassland or short mopane. A post-processing filter in conjunction with semi-supervised learning should be able to learn such high-level information and improve the classification results.

6.2.3 Disjointed Shortest Path Network

Studies presented in Section 5.3 show that generating a shortest path network is slow because the number of pairwise distances grows exponentially with the number of total samples. This research found that the majority of these pairwise relations are irrelevant to the classification process. Instead of building one large, fully connected network, future researchers could extend the findings of this dissertation to create disconnected shortest path networks that are spatially collocated and spectrally similar. These disjointed networks would preserve the framework by transferring knowledge learned from one region to the new region while avoiding the unnecessary calculations of the shortest path distances. Additional research should focus on

analyzing the appropriate threshold to define the spectral neighborhood and its impact on classification accuracies.

Given the reduced processing time, an ensemble of SkNN could be considered to reduce the standard deviation of classification accuracies. For instance, Section B.3 shows that bagging the labeled samples can reduce the variance of the final ensemble decision. Future studies that integrate the disjointed STP network and bagging could provide higher accuracies in the new region and mitigate the impact of a slow processing time.

Algorithms presented in this research can be extended for other applications. Similar advances in biomedical forensics and material science are providing similar information for identification of diseases and material characteristics. In all these applications, advanced methods are required to exploit the data rapidly, but users may not be algorithm experts.

Appendix A

Data

A.1 Kennedy Space Center, Florida

The NASA AVIRIS (Airborne Visible/Infrared Imaging Spectrometer) instrument acquired data over the Kennedy Space Center (KSC), Florida, on March 23, 1996. AVIRIS acquires data in 224 bands of 10 nm width with center wavelengths from 400 - 2500 nm. The KSC data, acquired from an altitude of approximately 20 km, have a spatial resolution of 18 m. After removing water absorption and low SNR bands, 176 bands were used for the analysis. Training data were selected using land cover maps derived from color infrared photography provided by the Kennedy Space Center and Landsat Thematic Mapper (TM) imagery. The vegetation classification scheme was developed by KSC personnel in an effort to define functional types that are discernible at the spatial resolution of Landsat and these AVIRIS data. Discrimination of land cover for this environment is difficult due to the similarity of spectral signatures for certain vegetation types. For classification purposes, 13 classes representing the various land cover types that occur in this environment were defined for the site. (Table A.1). Classes 4 and 6 represent mixed classes.

Ten randomly sampled partitions of the training data were sub-sampled such

Table A.1: Class Codes, Names, and Number of Training Samples for Kennedy Space Center AVIRIS

	Class	No. samples
1	Scrub	761 (14.6%)
2	Willow swamp	243 (4.66%)
3	Cabbage palm hammock	256 (4.92%)
4	Cabbage palm/oak hammock	252 (4.84%)
5	Slash pine	161 (3.07%)
6	Oak/broadleaf hammock	229 (4.38%)
7	Hardwood swamp	105 (2.00%)
8	Graminoid marsh	431 (8.27%)
9	Spartina marsh	520 (9.99%)
10	Cattail marsh	404 (7.76%)
11	Salt marsh	419 (8.04%)
12	Mud flats	503 (9.66%)
13	Water	927 (17.8%)

that 75% of the original data were used for training and 25% for testing. In order to investigate the impact of the quantity of training data on classifier performance, training data were then sub-sampled to obtain ten samples comprised of 50%, 30%, and 15% of the original training data. All classifiers were evaluated using the ten test samples composed of 25% of the original training data.

A.2 Okavango Delta, Botswana

The NASA EO-1 satellite acquired a sequence of data sets over the Okavango Delta, Botswana in 2001-2003. Preprocessing of the data was performed by the UT Center for Space Research to mitigate the effects of bad detectors, inter-detector miscalibration, and intermittent anomalies. Uncalibrated and noisy bands that cover water absorption features were removed, and the remaining 145 bands were included as candidate features: [10-55, 82-97, 102-119, 134-164, 187-220]. Multiple test sets were selected from this area because of the coverage of the Hyperion sensor and the

availability of multiple images covering the same area.

A.2.1 Fourteen Class Botswana Test and Spatially Disjoint Test Sets

The first Botswana data analyzed in this study was acquired on May 31, 2001. It consists of observations from 14 identified classes representing land cover types in seasonal swamps, occasional swamps, and drier woodlands located in the distal portion of the Delta. The class names and corresponding numbers of ground truth observations used in the experiments are listed in Table A.2. Classes 3 and 4 are both floodplain grasses that are seasonally inundated but differ in their hydroperiod (the amount of time inundated). Classes 9 through 11 represent different mixtures of acacia woodlands, shrublands, and grasslands and are named according to the dominant class. Training data were selected manually from the following sources: global positioning system located vegetation surveys, aerial photography from the Aquarap (2000) project, and 2.6-m resolution IKONOS multispectral imagery.

Similar to KSC dataset, ten randomly sampled partitions of the training data were sub-sampled such that 75% of the original data were used for training and 25% for testing. In order to investigate the impact of the quantity of training data on classifier performance, training data were then sub-sampled to obtain ten samples comprised of 50%, 30%, and 15% of the original training data. All classifiers were evaluated using the ten test samples composed of 25% of the original training data.

Spatially Disjoint (S. D.) Test Set

Due to the size of the image and the ground truths variety, additional labeled samples were selected from spatially disjoint locations for further testing. The individual class signatures are slightly different and are used to test the robustness of a classifier. The purpose for collecting the SD test data is explained in Section 2.3. The number

Table A.2: Class Codes, Names, and Number of Training Samples for Botswana Hyperion Data

	Class	Number of samples
1	Water	270 (8.31%)
2	Hippo grass	101 (3.09%)
3	Floodplain grasses1	251 (7.74%)
4	Floodplain grasses2	215 (6.63%)
5	Reeds	269 (8.27%)
6	Riparian	269 (8.27%)
7	Firescar	259 (7.98%)
8	Island interior	203 (6.26%)
9	Acacia woodlands	314 (9.67%)
10	Acacia shrublands	248 (7.65%)
11	Acacia grasslands	305 (9.38%)
12	Short mopane	181 (5.56%)
13	Mixed mopane	268 (8.27%)
14	Exposed soils	95 (2.92%)

of samples for each class is listed in Table A.3.

A.2.2 Nine Class Test and Spatially Disjoint Test Sets

Several overlapping classes used in the 14 classes test set were aggregated to form this 9 class test set to avoid confusion. For example, short mopane and mixed mopane are aggregated into short mopane. These samples were reselected to reflect the distribution of the area. The number of samples for each class is shown in Table A.4. Similar to the 14 classes test set, samples selected from spatially disjoint locations are also included for testing. Details are shown in Table A.5.

A.2.3 Edge Test Set

Because the training and test data are spatially collocated and selected from a relatively homogeneous area, an extended test set was also acquired and used to evaluate the generalization of these classifiers to other areas that are often on class bound-

Table A.3: Class Codes, Names, and Number of Spatially disjoint Test Samples for Botswana Hyperion Data

	Class	Number of S. D. samples
1	Water	126 (5.05%)
2	Hippo grass	162 (6.50%)
3	Floodplain grasses1	158 (6.34%)
4	Floodplain grasses2	165 (6.62%)
5	Reeds	168 (6.74%)
6	Riparian	211 (8.46%)
7	Firescar	176 (7.06%)
8	Island interior	154 (6.17%)
9	Acacia woodlands	151 (6.05%)
10	Acacia shrublands	190 (7.62%)
11	Acacia grasslands	358 (14.35%)
12	Short mopane	153 (6.13%)
13	Mixed mopane	233 (9.34%)
14	Exposed soils	89 (3.57%)

aries. Note that this extended data may have substantially different characteristics as it is taken from mixed locations. The purpose of having this data set is to investigate the capability of various methods that utilize contextual information in land cover classification. Hereafter, these data are referred to as the edge test data.

A.2.4 Multi-temporal Test Set

Solving a knowledge transfer problem involves using training models developed from data acquired earlier to classify newly obtained images. Thus, two additional Hyperion images acquired over the Okavango Delta, Botswana on June 16th and July 1st of 2001 were processed and additional labeled samples were selected for multi-temporal experiments. The original May 31st data set and two newly processed data sets were labeled as May, June and July data sets respectively. The relation of these three data sets are presented in Figure A.1. Details of June and July data sets are shown in Table A.7 and A.8 respectively.

Table A.4: Botswana Training Data: Individual Class

Class	Number of Pixels
Water	158
Primary Floodplain	228
Riparian	237
Firescar	178
Island Interior	183
Woodlands	199
Savanna	162
Short Mopane	124
Exposed Soils	111

Table A.5: Botswana Spatially Disjoint Test Data: Individual Class

Class	Number of Pixels
Water	139
Primary Floodplain	209
Riparian	211
Firescar	176
Island Interior	154
Woodlands	158
Savanna	168
Short Mopane	115
Exposed Soils	104

Table A.6: Botswana Edge Test Data: Individual Class

Class	Total Numbers
Water	145
Primary Floodplain	145
Riparian	140
Firescar	152
Island Interior	149
Woodlands	133
Savanna	141
Short Mopane	143
Exposed Soils	155

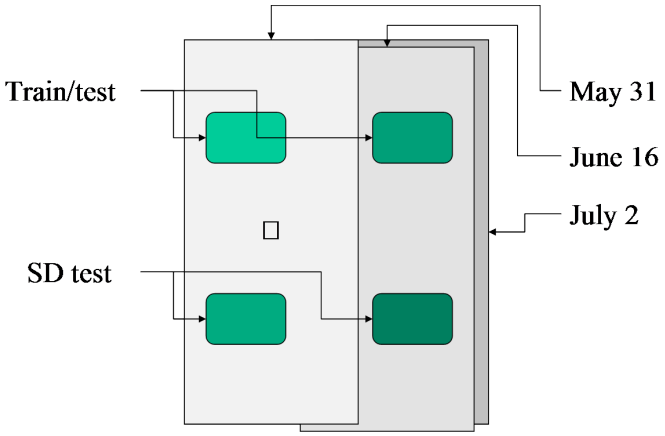

Figure A.1: Multi-Temporal Data

Table A.7: Botswana June Multi-temporal Test Data: Individual Class

Class	Total Numbers
Water	196
Primary Floodplain	192
Riparian	180
Firescar	196
Island Interior	200
Woodlands	220
Savanna	192
Short Mopane	168
Exposed Soils	156

Table A.8: Botswana July Multi-temporal Test Data: Individual Class

Class	Total Numbers
Water	188
Primary Floodplain	100
Riparian	164
Firescar	188
Island Interior	132
Woodlands	172
Savanna	172
Short Mopane	152
Exposed Soils	102

Appendix B

Ensemble Methods

Statistical classification of hyperspectral data is challenging because the inputs are high in dimension and represent multiple classes that are sometimes quite mixed, while the amount and quality of ground truth in the form of labeled data is typically limited. The resulting classifiers are often unstable and have poor generalization. This chapter investigates two approaches based on the concept of random forests of classifiers implemented within a binary hierarchical multi-classifier system, with the goal of achieving improved generalization of the classifier in analysis of hyperspectral data, particularly when the quantity of training data is limited. A new classifier is proposed which incorporates bagging of training samples and adaptive random subspace feature selection within a Binary Hierarchical Classifier (BHC), such that the number of features that is selected at each node of the tree is dependent on the quantity of associated training data. Results are compared to a random forest implementation based on the CART framework. For both methods, classification results obtained from experiments on data acquired by the NASA airborne AVIRIS instrument over the Kennedy Space Center, Florida, and by Hyperion on the NASA EO-1 satellite over the Okavango Delta of Botswana are superior to those from the original best basis BHC algorithm and a random subspace extension of the BHC.

B.1 Introduction

The increasing availability of data from hyperspectral sensors, particularly with the launch of the Hyperion instrument on the NASA EO-1 satellite, has generated tremendous interest in the remote sensing community. These instruments characterize spectral signatures with much greater detail than traditional multispectral sensors and thereby can potentially provide improved discrimination of targets [59]. However, hyperspectral data also present difficult challenges for supervised statistical classification, where labeled training data are used to estimate the parameters of the label-conditional probability density functions [51]. The dimensionality of the data is high ($\sim$200), there are often tens of classes C, and the quantity of training data is often small. Sample statistics of training data may also not be representative of the true probability distributions of the individual class signatures, particularly for remote, inaccessible areas where training data are logistically difficult and expensive to acquire. Generalization of the resulting classifiers is often poor, thereby resulting in poor quality mapping over extended areas.

Various approaches have been investigated to mitigate the impact of small sample sizes and high dimensionality, which are inherently coupled issues since the adequacy of a data sample depends on the data dimensionality, among other factors [66]. For example, regularization methods try to stabilize the covariance matrix by weighting the sample covariance matrix and a pooled covariance matrix or by shrinking the sample covariance matrix toward the identity matrix [73]. While this may reduce the variance of the parameter estimates, the bias of the estimates can increase dramatically. Alternatively, the input space can be transformed into a reduced feature space via feature selection [71] or feature extraction. Although these two approaches reduce the effect of the high dimensionality problem, feature selection methods are often trapped in a local optimal feature subset, while feature extraction methods lose the interpretability of the original features. Another way of

dealing with a small training set is to augment it with unlabeled data and then use semi-supervised learning techniques. These methods have been shown to enhance supervised classification [39, 6]. However, convergence of the updating scheme can be problematic, and it is affected by selection of the initial training samples and by outliers.

In analysis of hyperspectral data, Lee and Landgrebe proposed methods for feature extraction based on decision boundaries that maximize separation of data in multiple two-class problems [52]. These decision boundary feature extraction (DBFE) methods are often effective for two-class problems, but do not exploit correlation between sequential bands. Jia and Richards developed the Segmented Principal Components Transformation (SPCT) whereby the original bands are grouped into subsets of highly correlated adjacent bands to which the K-L transform is applied. The most significant principal components are then selected from each subset to yield a feature vector with reduced dimension [41]. The approach treats interband correlation globally and does not guarantee good discrimination capability because the PCT preserves variance in the data rather than maximizing discrimination between classes. Kumar *et al.* investigated band combining techniques, motivated by best-basis functions, as a means of feature extraction in a pairwise classifier framework [48]. Adjacent bands are selected for merging (alt. splitting) in a bottom-up (alt. top down) fashion using the product of a correlation measure and a Fisher discriminant. Morgan *et al.* [56] suggested a similar correlation-based band combining approach, in conjunction with a covariance shrinkage method, for both a top-down and bottom up hierarchical classifier to ameliorate the small training data problem.

The theory and practice of classifier ensembles also provide ways of alleviating sample size and high dimensionality concerns [50]. Bagging involves bootstrapped sampling of the original data, generating a classifier specific to each sample, and

then averaging the classifier outputs [8]. This method takes advantage of data reuse, but when the training data set in the (sub-)sample is very small, the potential for improved diversity and reduced impact of outliers is offset by degradation in individual classifier performance [77]. Boosting also combines weak individual classifiers to develop an improved classifier, but by re-weighting training data to increase sensitivity to incorrectly classified training observations. While boosting can improve performance for large training samples, it is not useful for small sample problems, particularly in the presence of outliers. When the input space is large, random subspace (RS) feature selection can potentially provide improved classifier diversity, while stabilizing parameter estimates, by randomly reducing the number of inputs to each classifier in the ensemble and constructing multiple classifiers in the resulting random input space [37, 72]. The method is potentially attractive for problems with redundant input features (e.g., hyperspectral data) and when outliers exist in the training data. Recently, approaches referred to as "random forests of classifiers" have been proposed. These involve developing multiple trees from randomly sampled subspaces of input features, then combining the resulting outputs via voting or a maximum *a posteriori* rule [9]. These methods typically achieve superior generalization for small training samples, but are computationally intensive.

Land cover classification problems usually involve a large number of classes, i.e. the output space is large. Output decomposition using binary classifiers in a multi-classifier framework has been shown to be more successful than traditional 1-of-C classifiers for many problems involving large output spaces [29]. Decomposition methods using pairwise classifiers, error correcting output codes (ECOC) [24], and binary decision trees have all been investigated in this context (see [56] for an overview). Pairwise classifiers develop a separate classifier for each pair of classes, thereby resulting in $O\left(C^2\right)$ classifiers which must be combined to determine the final class label. These methods often yield simple classifiers with excellent dis-

crimination for specific pairs, but are generally inefficient for problems with a large number of output classes. In the ECOC, a C-class problem is decomposed into $\hat{C}$ binary problems, whereby the original class is then encoded into a $\hat{C}$ binary vector of a coding matrix. It has been shown that the ECOC method yields robust, stable classifiers. However, since the code matrix design is not based on the characteristics of the classes it represents, interpretability of the classifier is limited.

Binary trees, which often provide an attractive approach for decomposing large output space problems, can be constructed using a variety of splitting functions involving single or multiple features and output classes. To address the high dimensional output problem while exploiting the affinity for spectrally similar classes, Kumar *et al.* proposed a Binary Hierarchical Classifier (BHC) [49] to decompose a $(C > 2)$-class problem into a binary hierarchy of $(C - 1)$ simpler 2-class problems that can be solved using a corresponding hierarchy of classifiers, each based on a simple linear discriminant. The method was extended by Morgan *et al.* [56] for small training samples using an adaptive best basis BHC, which exploits the class specific correlation structure between sequential bands of hyperspectral data and utilizes an adaptive regularization approach to stabilize covariance estimates. An adaptive random subspace feature selection approach was also investigated within the BHC framework (RS-BHC) as a means of improving classifier performance when the number of training samples is extremely small [17].

In this chapter, we investigate a random forest of binary classifiers as a means of increasing diversity of hierarchical classifiers. We evaluate the results obtained for trees produced by our BHC classifier and the original CART-based random forest method [9]. For the BHC, the goal is to exploit the advantages of natural class affinity while improving generalization in classification of hyperspectral data when the number of training samples is small. The CART-based approach is not directly affected by small sample size statistics and potentially provides greater diversity

within the forest, but typically produces trees of enormous size if the output space is large. The chapter is organized as follows: the best basis (BB-BHC), random subspace (RS-BHC), and random forest (RF-BHC) implementations of the BHC method and the CART-based framework (RF-CART) are all described in Section II; classification results using the random forest approaches obtained for data acquired by AVIRIS (Airborne Visible/Infrared Imaging Spectrometer) over the Kennedy Space Center, Florida, and EO-1 Hyperion over the Okavango Delta, Botswana, are presented in Section III and compared to those obtained from the BB-BHC, RS-BHC. Results from all the methods are evaluated, and new directions for future work are suggested in Section IV.

B.2 Random Forest Binary Hierarchical Classification Method

The top-down Binary Hierarchical Classifier (BHC) framework recursively decomposes a C-class problem into $C-1$ two-(meta)class problems via a deterministic simulated annealing method [49]. The root classifier tries to optimally partition the original set of classes into two disjoint meta-classes while simultaneously determining the Fisher discriminant that separates these two subsets. This procedure is recursed, i.e., the meta-class Ω_n at node n is partitioned into two meta-classes $(\Omega_{2n}, \Omega_{2n+1})$, until the original C classes are obtained at the leaves. The tree structure, as shown in Figure B.1, allows the more natural, easier discriminations to be accomplished earlier. Fewer classes are involved in the partitioning at lower levels of the BHC hierarchy. Thus, while the classification task typically becomes simpler, the number of relevant training samples also decreases. The BB-BHC ameliorates this effect by utilizing an ancestor covariance matrix while exploiting the inter-band serial correlation through an adaptive, class dependent, band aggregation process

[56]. A band combining step is performed on highly correlated, spectrally adjacent bands prior to the partitioning of meta-classes, thereby reducing the number of inputs relative to the number of training data points. Bands are aggregated until a user defined ratio, R, between the number of training samples for the respective (meta)classes and input dimension is achieved. Typically, R is selected to be at least 5.

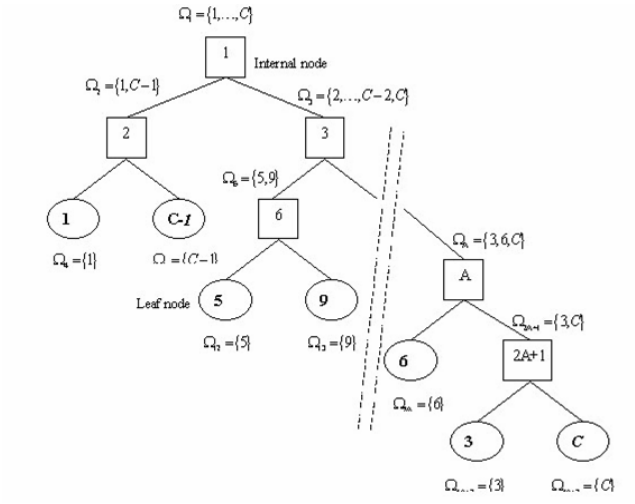

Figure B.1: Binary Hierarchical Classifier Framework for Solving a C-class Problem

The BHC method is extended in the RS-BHC approach by utilizing the random subspace method as a post-processing stage to tree construction, with the goal of reducing the number of inputs while refining decision boundaries [58]. The BB-BHC method is used to first construct the hierarchy, then random subspace sampling is performed at each node of the tree where the criterion for R is not satisfied. For each (meta) class m with n_m vector-valued observations, $X_m = (X_1, \ldots X_{nm})$, a subset of elements of $X_i = (X_{i1}, \ldots, X_{ik})$ with dimension $p_m = n_m/R < k$ is then randomly selected from the k-dimensional set of features. The resulting modified training set $X_m^r = (X_1^r, \ldots X_{nm}^r)$ consists of observation vectors, $X_i^r = (X_{i1}^r, \ldots, X_{ik}^r)$, where the same subset of features is selected for each

element $X_i^r \in X^r (i = 1, \ldots, n_m)$. The number of random subspaces selected at each node is $N_s = (k/p_m) \times F$ where the value of F is a user supplied input. A discriminant vector is constructed for each random subspace, and N_s the vectors are combined at each node of the hierarchy via majority voting. Our empirical evidence indicates that good results are typically achieved for $2 < F < 4$, which provides adequate coverage of the feature space. Improvement in classification accuracy is not significant for $F > 4$.

The random forest implementation of the BHC (RF-BHC) extends the RS-BHC by incorporating random subspace feature selection in the actual development of the tree. This is particularly advantageous as random subspace sampling is performed by the RS-BHC only at nodes where the ratio, R, is not exceeded. Thus, sub-sampling of the input features typically occurs only at lower levels of the tree, thereby limiting diversity. For moderate sized training samples, bagging can increase diversity of the multiclassifier system, so a bootstrap sample of observations is selected for each tree in the RF-BHC. At each metaclass node m, a random subspace of features of dimension $p_m = \min(p_m, N_f)$ is selected to determine the decision boundary for the classifier at that node, where N_f is a user selected input. To guarantee greater diversity, we choose $N_f \ll k$. The tree is then developed using the resulting set of features selected at each node. The process is repeated to grow a forest of identically, independently distributed random vectors associated with the individual trees.

The fundamental difference between BHC and other decision trees is that the former focuses on decomposing the output space; partitioning of the input space occurs as a consequence. Both RF-BHC and RF-CART use the random forest ensemble method to increase the diversity of each base learning module, then combine results of the individual modules (trees). While Breiman's CART-based random forest follows a typical binary divide-and-conquer hierarchical scheme, it differs from

the BHC in the base learning module. The BHC uses the GAMLS [47] algorithm to split each node into meta-classes which are separated by the maximum Fisher distance. Using a sequence of binary tests, CART seeks the split that maximizes the reduction of the impurity of the parent nodes and its two child nodes as measured by the Gini index [10]. The most discriminating feature is selected to perform the split. Used in the random forest context, a random subspace of the original k features is selected at each node of the tree, and the most discriminating feature is then selected. Further, unlike the actual CART method, the RF-CART approach does not perform pruning of nodes as pruning reduces diversity of trees in the forest. Analogous to the RF-BHC, each tree is grown using a bootstrap sample of the training set.

B.3 Results

Hyperspectral data from two sources were analyzed in this chapter:

1. Kennedy Space Center, Florida: The NASA AVIRIS (Airborne Visible/Infrared Imaging Spectrometer) instrument acquired data over the Kennedy Space Center (KSC), Florida, on March 23, 1996. AVIRIS acquires data in 224 bands of 10 nm width with center wavelengths from 400 - 2500 nm. The KSC data, acquired from an altitude of approximately 20 km, have a spatial resolution of 18 m. After removing water absorption and low SNR bands, 176 bands were used for the analysis. Training data were selected using land cover maps derived from color infrared photography provided by the Kennedy Space Center and Landsat Thematic Mapper (TM) imagery. The vegetation classification scheme was developed by KSC personnel in an effort to define functional types that are discernable at the spatial resolution of Landsat and these AVIRIS data. Discrimination of land cover for this environment is difficult due to the similarity of spectral signatures for certain vegetation types. For classification

purposes, 13 classes representing the various land cover types that occur in this environment were defined for the site. (Table A.1). Classes 4, and 6 represent mixed classes.

2. Okavango Delta, Botswana: The NASA EO-1 satellite acquired a sequence of data over the Okavango Delta, Botswana in 2001-2004. The Hyperion sensor on EO-1 acquires data at 30 m pixel resolution over a 7.7 km strip in 242 bands covering the 400-2500 nm portion of the spectrum in 10 nm windows. Preprocessing of the data was performed by the UT Center for Space Research to mitigate the effects of bad detectors, inter-detector miscalibration, and intermittent anomalies. Uncalibrated and noisy bands that cover water absorption features were removed, and the remaining 145 bands were included as candidate features: [10-55, 82-97, 102-119, 134-164, 187-220]. The data analyzed in this study, acquired May 31, 2001, consist of observations from 14 identified classes representing the land cover types in seasonal swamps, occasional swamps, and drier woodlands located in the distal portion of the Delta [22]. These classes were chosen to reflect the impact of flooding on vegetation in the study area. The class names and corresponding numbers of ground truth observations used in the experiments are listed in Table A.2 and A.3. Classes 3 and 4 are both floodplain grasses that are seasonally inundated, but differ in their hydroperiod (the amount of time inundated). Classes 9, 10, and 11 represent different mixtures of acacia woodlands, shrublands, and grasslands and are named according to the dominant class. Training data were selected manually using a combination of GPS located vegetation surveys, aerial photography from the Aquarap (2000) project, and 2.6 m resolution IKONOS multispectral imagery. The class priors for both data sets, as indicated by the labeled data, are only moderately skewed. For simplicity, we assume the class priors to be equal while developing the BHC classifier. This assumption shall

be reconsidered later.

For both data sets, ten randomly sampled partitions of the training data were sub-sampled such that 75% of the original data were used for training and 25% for testing. In order to investigate the impact of the quantity of training data on classifier performance, these training data were then sub-sampled to obtain ten samples comprised of 50%, 30%, and 15% of the original training data. All classifiers were evaluated using the ten test samples composed of 25% of the original training data.

Experiments were performed using the BB-BHC, RS-BHC, RF-BHC, and RF-CART. Although authors recommend various values for the dimension of the random subspace and the number of trees in a random forest, there do not appear to have been any systematic studies of the issue to date. In the results reported here, the ratio R was set at 5, and the value of F was 4 for the RS-BHC method. In our experiments, the dimension of the random subspace was determined adaptively in the BHC, but was always selected such that the value of R was at least 5. For the RF-BHC, the value of N_f was selected to be 20. In order to have somewhat comparable inputs, 20 input features were randomly selected in the RF-CART method. 100 trees were grown for each experiment as our sensitivity studies showed that larger forests did not provide improved results for these data sets.

B.3.1 Original Training and Test Areas

Kennedy Space Center, Florida: The true color image shown in Figure B.2, along with the classification results obtained from the RF-BHC in Figure B.3, shows the spatial distribution of classes and training sites over the 614 x 512 pixel study area.

Average classification accuracies for test data and associated standard deviations for the 10 experiments conducted with each classifier are plotted in Figures

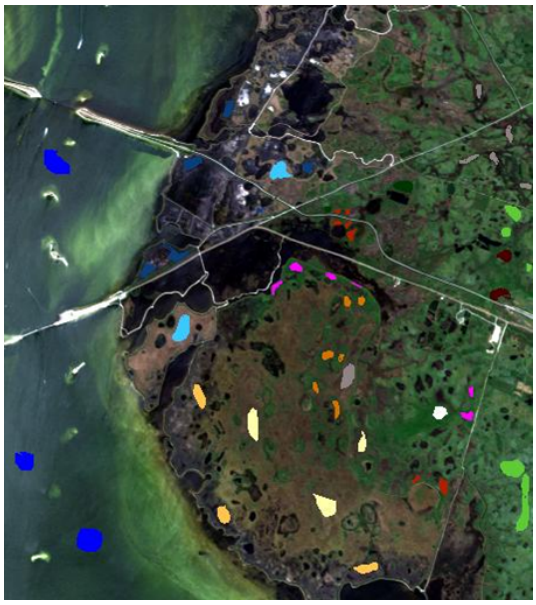

Figure B.2: AVIRIS Data, (Bands 31, 21, 11) Acquired over KSC, Training Sites Overlaid

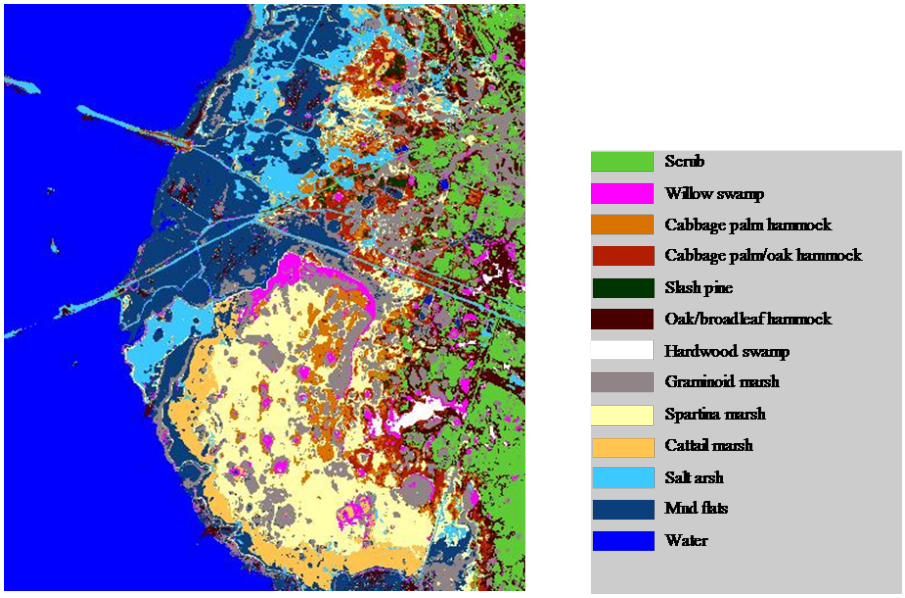

Figure B.3: Classified Image of KSC AVIRIS Data using RF-BHC Classifier

B.4 and B.5. The overall trends in accuracies relative to the quantity of training data are similar for all methods when applied to the test data set. At the 75% sampling rate, the accuracies of all methods are nearly the same, although the RF-BHC yields somewhat higher accuracies than the other methods. The results obtained by the BB-BHC and RF-CART methods are very similar over all sampling rates, with RF-CART yielding slightly lower accuracies. These methods consistently produced the lowest overall average accuracies. The RS-BHC yielded approximately the same accuracies as the BB-BHC at the 75% sampling rate, but improved relative to the BB-BHC and the RF-CART approach at lower sampling rates. For the BHC-based methods, this appears to demonstrate the value of reduced redundancy in the input space and improvements achieved by better tuning of the decision boundaries, even though the tree structure is identical to the BB-BHC and random sampling of the feature space is not required until lower levels of the tree (particularly for the higher training data fractions). Results were also obtained using the original RS-BHC and best basis aggregated data. The BB-RS-BHC consistently yielded slightly lower accuracies than the RS-BHC because of the reduced diversity of trees, but results were not statistically different and are not reported here. The overall average accuracy of the RF-BHC is consistently the highest and improves relative to other BHC methods and RF-CART as the fraction of training data is reduced.

The RF-BHC is also the most stable method over all training fractions, as measured by the standard deviation of the accuracies. The standard deviations of the accuracies of the random subspace based methods appear to benefit from the diversity of the input space. The standard deviation of the accuracies obtained by the BB-BHC increased dramatically at the 30% sampling rate because it was necessary to aggregate a large number of bands to satisfy the ratio R. The problem, which is manifested both in the tree building and in the decision boundary of the BB-BHC, is offset in the determination of the RS-BHC decision boundary. It should

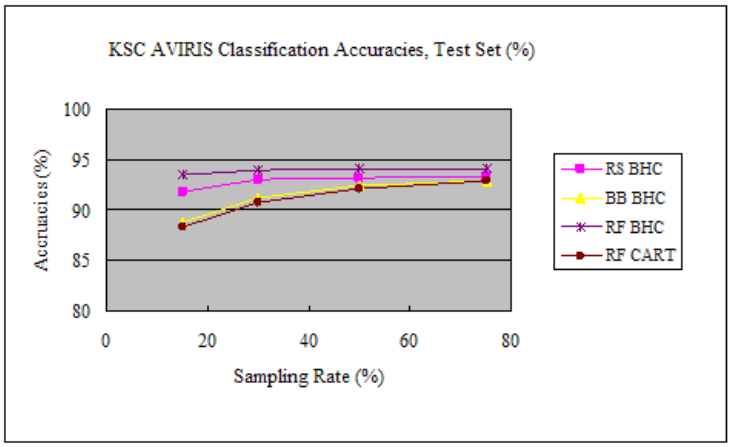

Figure B.4: Ave of Classification Accuracies for AVIRIS Test Data

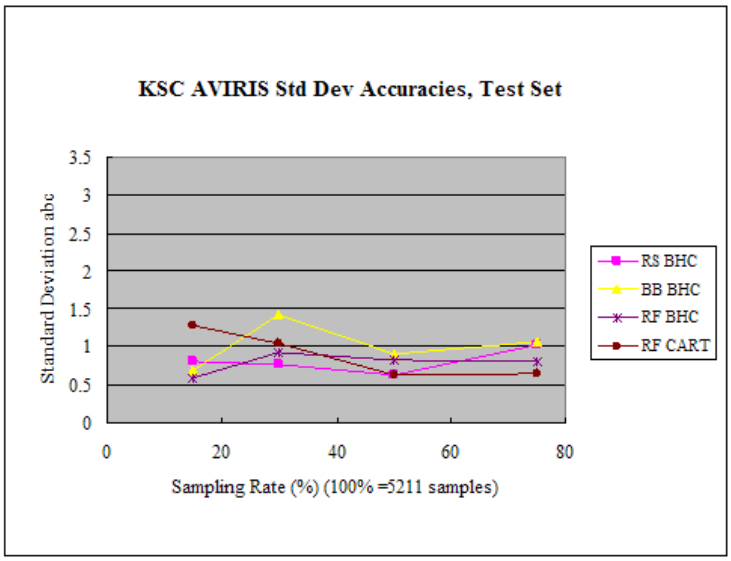

Figure B.5: Std. Dev. of Classification Accuracies for AVIRIS Test Data

be noted that although the standard deviation of BB-BHC decreases at the 15% sampling rate, the associated average classification accuracy is also poor, further demonstrating it is uniformly inferior at low sampling rates. The reduced accuracy of RF-CART at low sampling rates, relative to the BHC based random forest methods, is attributed to the value of the inherent exploitation of class affinities by the BHC approaches. Further, although the standard deviation of the accuracies for the RF-CART approach is low for high sampling rates, it increases consistently as the sampling rate of the training data is reduced, likely because the discrimination capability of the single best feature within a small random sample of inputs may be quite variable. Further, the benefits of bagging the training sample occur at the higher sampling rates for both the RF-BHC and RF-CART methods.

Okavango Delta, Botswana: The RGB image in Figure B.6 and the classification results obtained by the RF-BHC in Figure B.7 show that the spatial distribution of classes is extremely complex over this 256 x 1476 pixel area. Using the same random sampling strategy as for the KSC data, results were obtained at each percentage for all four classifiers.

Plots of classification accuracies at the various sampling rates are shown in Figures B.8 and B.9. The overall trends in accuracies relative to the fraction of training data are similar those of the KSC AVIRIS test data. Among the classifiers, RF-BHC yielded the highest classification accuracy on all training sample fractions, and performance degraded only slightly at lower sampling rates. The standard deviations of the accuracies obtained using the RF-BHC are low and remain nearly constant over the various sampling rates. At the 30% sampling rate, the standard deviations of the accuracies yielded by the BB-BHC and RS-BHC are substantially higher. For the BB-BHC, this again appears to be due to the amount of band aggregation required to achieve the ratio R. Unlike the KSC case, the RS-BHC is apparently unable to mitigate problems associated with band aggregation during

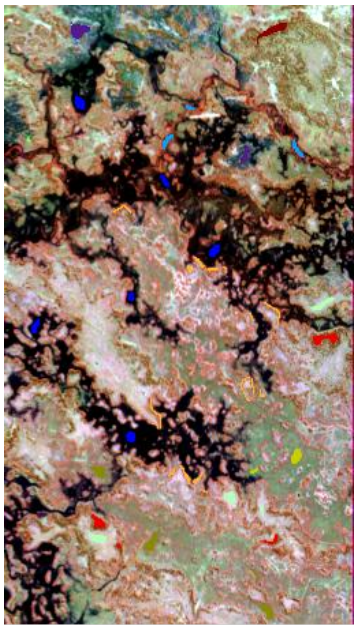

Figure B.6: Hyperion Data, (Bands 51, 149, 31) Acquired ver Okavango Delta, Training Sites Overlaid

the tree construction phase for the Okavango data.

The difference in results produced by the RF-BHC and RF-CART methods was unexpected since both utilize an ensemble of 100 trees to build a stronger classifier. Previous research by Tumer and Ghosh [14] indicated that the accuracy of an ensemble method relies on the diversity of the base classifier. To investigate the performance of the individual trees, we further analyzed the performance of both random forest methods at the 75% sampling rate. The average accuracy over the set of individual trees developed by the RF-BHC is 89.2%, and the standard deviation is 1.3. The overall accuracy for the RF-BHC, which is determined by simple voting, increased by 5.7% to 94.9%. For the RF-CART method, the average accuracy obtained using 100 trees is 84.2%, with standard deviation 1.3. The ensemble of these 100 trees, using simple voting utilized in the original Brieman random forest, yielded a 7.8% increase to 92%. For this type of classification problem, it appears that the

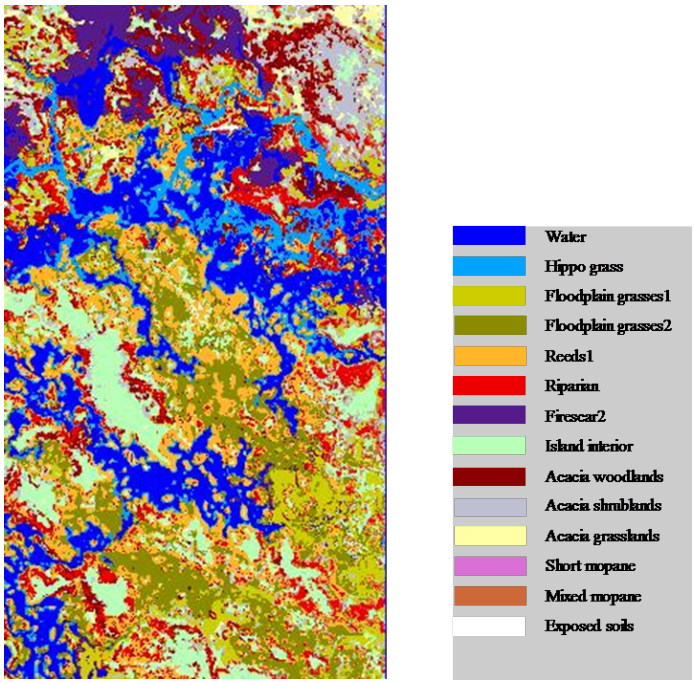

Figure B.7: Classified Image of Hyperion Data over Okavango Delta using RF-BHC Classifier

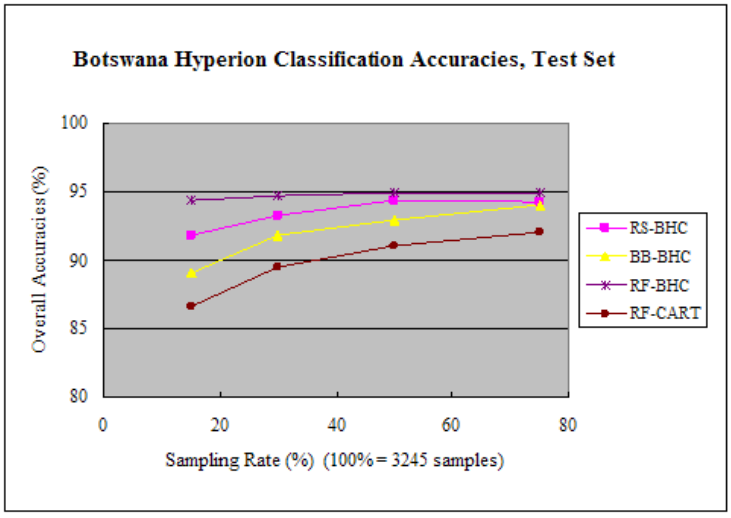

Figure B.8: Ave. of Classification Accuracies for Hyperion Test Sets

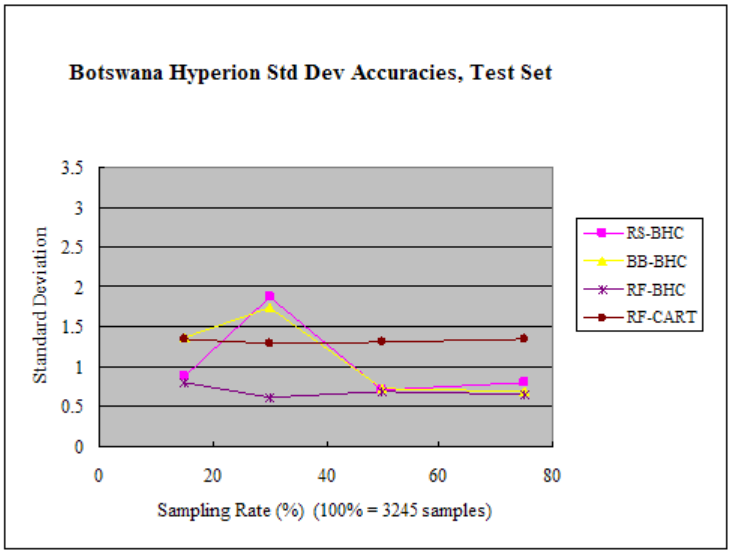

Figure B.9: Std. Dev. of Classification Accuracies for Hyperion Test Sets

BHC is a better base classifier than CART, although CART realizes substantial improvement when trees are combined.

B.3.2 Generalization to Spatially Disjoint Areas

Traditionally, the training and test data are spatially co-located and can thus be assumed to be samples from the same distribution. In practice, however, it is also useful to estimate how a classifier will perform in areas that are somewhat different, in order to indicate how much additional data labeling and retraining is needed to make the model applicable to much larger areas. With this goal in mind, a "spatially disjoint test" set was also acquired from a geographically separate location at the Botswana site and used to evaluate the classifiers developed previously.

These spatially disjoint data have somewhat different characteristics from the training/test data, so the performance of all classifiers is reduced, as expected. (See Figures B.10 and B.11.)

Still, as with the test data, the BB-BHC yielded the lowest overall average

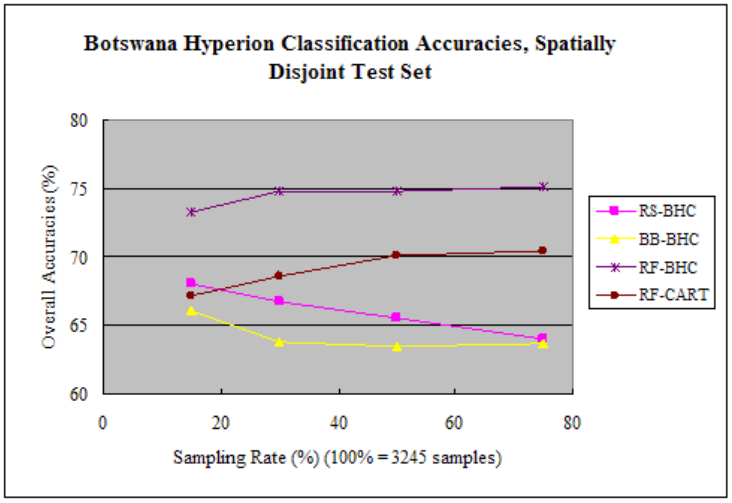

Figure B.10: Ave.of Classification Accuracies for Hyperion Spatially Disjoint Test Sets

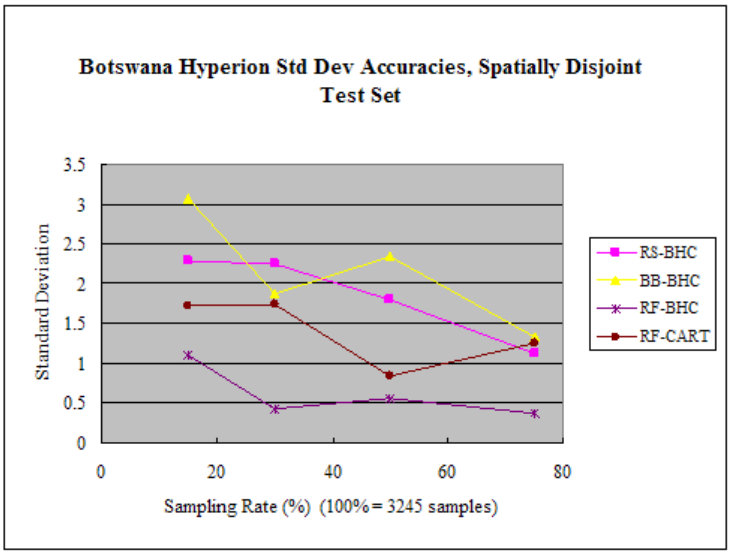

Figure B.11: Std. Dev. of Classification Accuracies for Hyperion Spatially Disjoint Test Sets

accuracy at all sampling rates. The incremental improvement in average accuracy achieved by the random subspace method increases with reduced sampling rates, but is not statistically significant as the standard deviations of the accuracies also increase substantially with lower sampling rates. The RF-BHC implementation and the RF-CART method yielded higher accuracies for the spatially disjoint test data at all sampling rates than both the BB-BHC and the RS-BHC, thereby demonstrating the greater generalization of these approaches. Similar to results from the test data, the RF-BHC consistently produced the highest overall average accuracies for the spatially disjoint test set, indicating the value of exploiting class affinity, coupled with the increased diversity of trees achieved by forcing random sampling of the input space at all nodes. The RF-CART method also achieved good generalization, as indicated by its performance on these spatially disjoint test data, although results for the original test set were inferior to the other methods. This is attributed both to the diversity that it achieves and its reduced dependence on the training sample statistics. Similar to the test data, the performance of both the RF-BHC and RF-CART methods was further investigated for the 100 trees obtained from the Hyperion spatially disjoint test set at the 75% sampling rate. The average classification accuracy over the set of individual RF-BHC trees is 68.2%, and the standard deviation is 2.9. The ensemble random forest result using simple voting is 75.2%, an increase of 7%. For RF-CART the values are 60.8% and 2.4, respectively. The classification accuracy increased by 9.6% to 70.4% when the 100 trees were combined using simple voting.

Since RF-BHC and RF-CART use the same random forest framework, their differences lie both in the tree construction and the underlying classifier. For the remotely sensed data in this study, the BHC exploits class affinity, while the good performance of the CART-like method on the spatially disjoint test set suggests that it provides more diversity. To further investigate this issue, we calculated the

entropy, a non-pairwise diversity measure [50], of trees obtained from both RF-BHC and RF-CART (Table B.12). The results indicate that RF-CART method produces more diverse trees than RF-BHC at all four sampling rates. RF-CART achieves an 11.2% increase in accuracies via the ensemble, while RF-BHC results improve only 7.7%, thereby reinforcing the idea that ensemble methods benefit more from combining diverse classifiers. Further, as the sampling rate increases, the diversity of RF-BHC trees decreases. Under the same situation, however, the diversity of the RF-CART forest remains comparatively consistent. This means that the RF-BHC inputs become more homogeneous as the number of samples increases, while RF-CART does not follow the same trend. Overall, the advantages of an ensemble approach are clear as the RS-BHC used only one tree structure rather than an ensemble of potentially different trees, which significantly reduced the generalization of its classification accuracies on the spatially disjoint test set.

Methods \ Sampling Rate		15%	30%	50%	75%
RF-BHC	Average	0.440	0.383	0.345	0.326
	Std	0.013	0.008	0.006	0.007
RF-CART	Average	0.516	0.476	0.460	0.460
	Std	0.024	0.020	0.011	0.010

Figure B.12: Entropy-based Diversity of Ensemble Members Observed for the Spatially Disjoint Botswana Hyperion Data at Different Sampling Rates

The differences between the overall accuracies for the spatially disjoint test set and those for the test set are quite remarkable. As noted earlier, the test data are spatially co-located with the training data, whereas the spatially disjoint test set is not. Clearly, either the class priors or the class conditional feature distributions (or both) are substantially different, at least for some classes in the more remote area. This motivated us to further investigate class specific results. Class dependent accuracies for the Hyperion test and spatially disjoint test sets are provided in Figures B.13 and B.14, and the detailed confusion matrix for the RF-BHC is contained

in Table B.15. Results in Table 2 indicate that the priors were indeed somewhat different. In particular, there were relatively more samples of Classes 2 and 11, and less of Classes 1 and 9. However, while false negative errors increase for Classes 2 and 11, there is no overall clear trend. For example, classification accuracies for Class 1 (water) which is spectrally quite distinct, are unaffected by the change of priors. Moreover, several class accuracies are now much lower than 80%, while others are almost unaffected. This leads us to believe that change in class-conditional distributions in certain classes that are spectrally quite similar is the main cause of the marked degradation in their classification accuracies. In particular, the overall classification accuracies of RS-BHC, RF-BHC and RF-CART methods are strongly influenced by the performance of Classes 2 and 11. Class 2, hippo grass, which grows within the river channels, has a small training sample and is spectrally similar to water as many pixels are mixed with water. Class 11, acacia grasslands, is a mixed class that is most often confused with other grasses or acacia shrubs, which is also a mixed class.

Using Figures B.16 and B.17, we can also compare the class specific accuracies for the spatially disjoint test set for the RS-BHC, RF-BHC and RF-CART approaches. Consistent with the overall accuracies, the performance of the RF-BHC is generally better than RF-CART method at both the 75% and 15% sampling rates. Similarly, the RS-BHC yields consistently lower accuracies, particularly for Classes 2 and 11. Although higher classification accuracies were achieved for Class 2 by RF-CART than the two BHC methods, it is not statistically significant as the standard deviation of the average sample accuracy is more than 12.

In comparing the overall computational requirements of the BHC-based and RF-CART methods, there are several tradeoffs. BHC-based trees always solve C-1 binary problems. At the 75% sampling rate, the average CART decision tree for Botswana Hyperion data contained 326 nodes (std. dev.= 16). For this 14 class

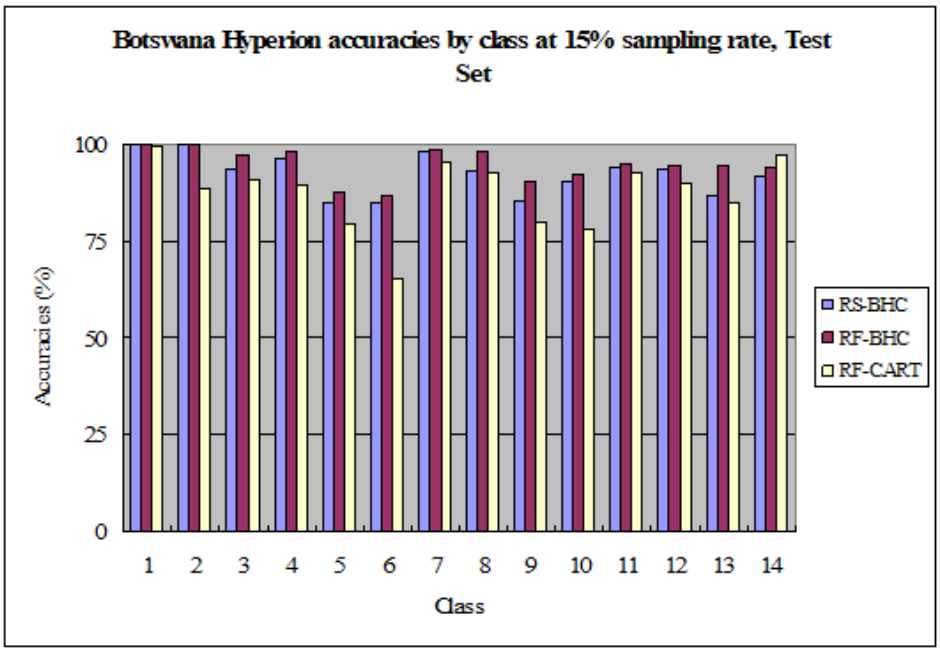

Figure B.13: Class Dependent Accuracies for Hyperion Test Set at 15% Sampling Rate

problem, the BHC tree had only 27 (1+13*2) nodes. For the same experiment, the CPU time for the RF-CART method was 8m 42s, while it was 1h 4m 4s for the RF-BHC. Both experiments were performed on a 3GHz Pentium 4 CPU machine. The RF-BHC required more CPU time than the RF-CART method because GAMLS is a deterministic simulated annealing algorithm. It should be noted that while neither algorithm was coded as an operational method, average timing results reflect their relative computational requirements.

B.4 Conclusion

The primary purpose of the study was to investigate the performance of random feature subset selection methods in terms of generalization. The secondary goal was to investigate the performance of the methods when applied to data acquired

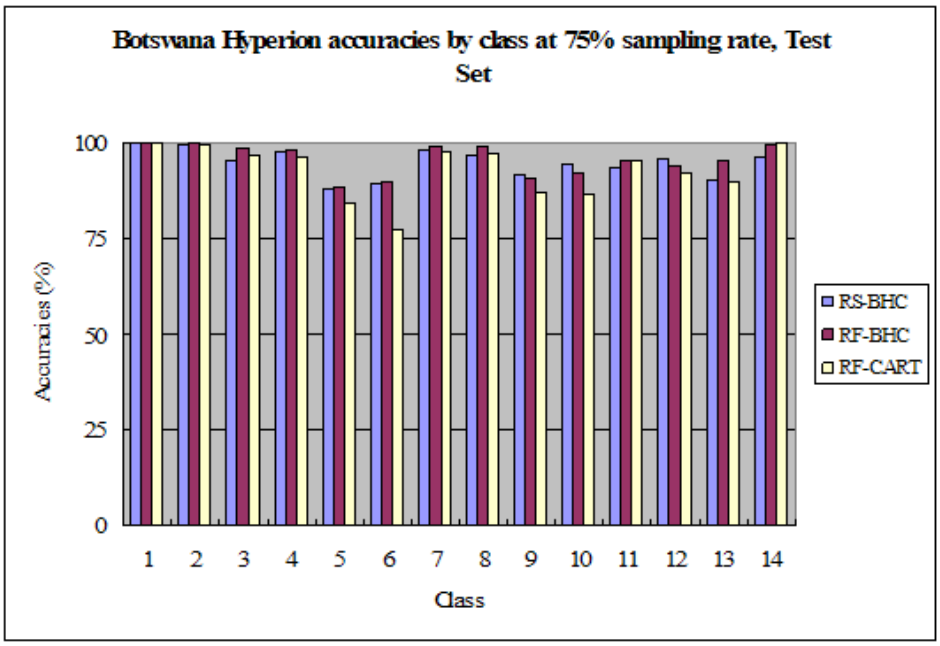

Figure B.14: Class Dependent Accuracies for Hyperion Test Set at 75% Sampling Rate

Classified\Actual (col)	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*
1	126	44	0	0	0	0	1	0	0	0	0	0	0	0
2	0	84	0	0	4	0	0	0	0	0	0	0	0	0
3	0	0	139	18	3	0	5	0	1	0	0	0	0	0
4	0	0	0	144	4	2	0	0	0	0	0	0	0	0
5	0	0	0	2	96	23	0	0	0	0	0	0	0	0
6	0	15	0	0	16	149	0	0	30	0	0	0	2	0
7	0	2	0	0	0	0	160	0	0	0	0	0	0	0
8	0	0	0	0	12	0	0	150	0	0	92	1	0	0
9	0	17	4	1	4	36	0	0	115	0	0	0	60	0
10	0	0	15	0	11	0	7	0	1	172	38	0	6	0
11	0	0	0	0	0	0	0	0	0	17	208	0	0	0
12	0	0	0	0	7	0	0	0	0	0	1	129	2	0
13	0	0	0	0	9	1	3	0	4	1	6	23	163	0
14	0	0	0	0	2	0	0	4	0	0	13	0	0	89
Total	126	162	158	165	168	211	176	154	151	190	358	153	233	89
Class accuracies	100	51.9	88	87.3	57.1	70.6	90.9	97.4	76.2	90.5	58.1	84.3	70	100

Figure B.15: Confusion Matrix for Hyperion Spatially Disjoint Test Set at 75% Sampling Rate, RF-BHC Classifier

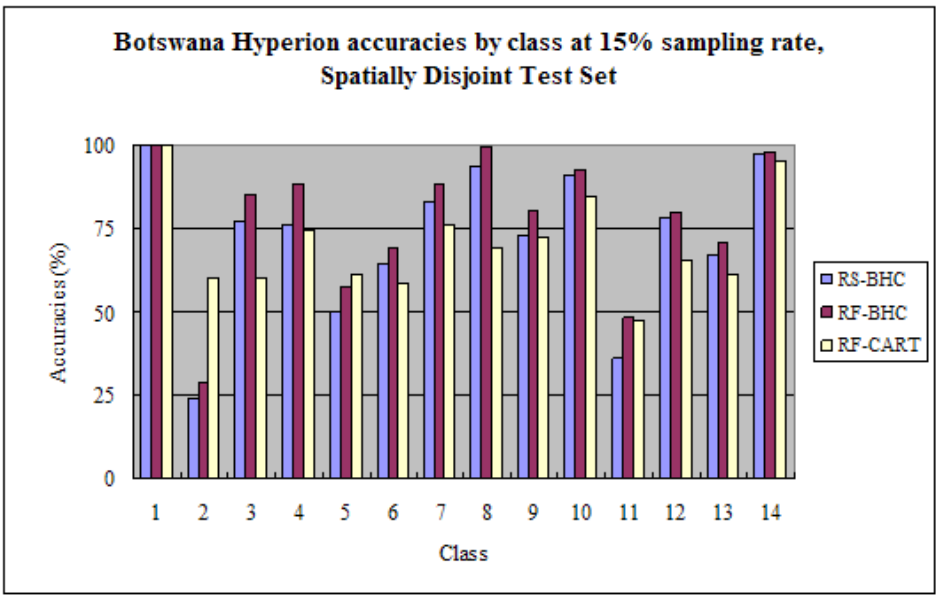

Figure B.16: Class Dependent Accuracies for Hyperion Spatially Disjoint Test Set at 15% Sampling Rate

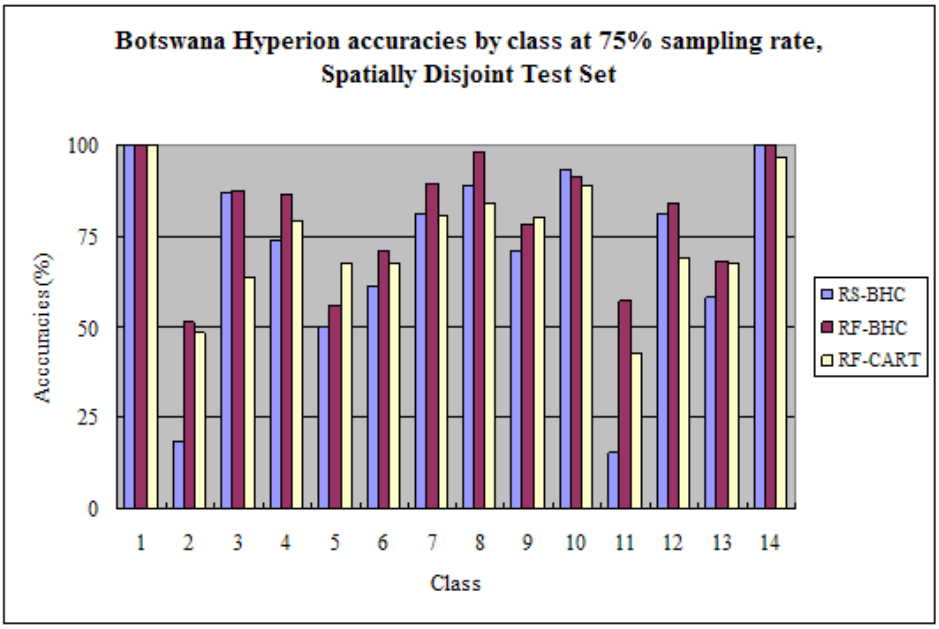

Figure B.17: Class Dependent Accuracies for Hyperion Spatially Disjoint Test Set at 15% Sampling Rate

by Hyperion data, which have low SNR. The performance of an implementation which focused on tuning decision boundaries of the BHC and that of two random forest approaches was investigated. Classification accuracies achieved by ensemble methods rely heavily on achieving diversity within the ensemble. The conflicting effects of improved SNR and reduced spectral resolution from band aggregation appear to be positively complemented by the improved diversity achieved by the RS-BHC through random sampling of the original features. We also noted that the change in classification accuracies achieved by using a forest rather than a single tree indicates that the RF-CART method actually achieves greater incremental benefit from the ensemble than the RF-BHC. Thus, the ensemble both exploits the greater diversity provided by the single feature splits and mitigates the potential impact of selecting features that are redundant or have poor discrimination capability.

A critical characteristic of the BHC is that it exploits the natural groupings of similar classes, which often occur in remotely sensed data acquired over natural landscapes. This provides a natural hierarchy which is often well handled by the simple Fisher discriminant. The random forest methods all yielded superior results for both test and spatially disjoint test data at our two study sites, with the improvement being greater for the spatially disjoint test set, thereby indicating improved generalization to extended areas. For these data, RF-BHC produced stable results over all sampling rates. Additional study is required to better characterize this issue. In this context, elimination of irrelevant and possibly redundant input features should also be considered in the RF-BHC. Other classifiers, such as the ECOC and SVM, should also be investigated within the RF-BHC framework. Overall, the RF-BHC methods appear to be quite promising in terms of generalization, but should be applied to many more data sets with different characteristics in order to better assess their overall performance. Also, much work remains to be done on determining how to improve performance on extended areas represented

by the spatially disjoint data set, especially since both the class mixtures and class conditional spectral properties can change in such situations. If this problem can be solved, then one can more confidently label much larger regions than those directly described by the available labeled data. For mixed classes, the issue may be mitigated in some cases by determining relative abundances of component classes via unmixing of hyperspectral data, if representative signatures of pure classes can be obtained [7, 2]. Approaches for representing spatially non-stationary spectral signatures may also be appropriate.

Bibliography

[1] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis.* Wiley, 2003.

[2] G. P. Asner and K. B. Heidebrecht. Spectral unmixing of vegetation, soil and dry carbon in arid regions: Comparing multi-spectral and hyperspectral observations. *Int. J. Remote Sens.*, 23:3939 – 3958, 2002.

[3] C. M. Bachmann, T. L. Ainsworth, and R. A. Fusina. Exploiting manifold geometry in hyperspectral imagery. *IEEE Trans. Geosci. and Remote Sens.*, 43(3):441–454, Mar 2005.

[4] C. M. Bachmann, T. L. Ainsworth, and R. A. Fusina. Improved manifold coordinate representations of large-scale hyperspectral scenes. *IEEE Trans. on Geosci. and Remote Sens.*, 44(10):2786 – 2803, October 2006.

[5] J. Besag. On the statistical analysis of dirty pictures. *Journal of Royal Statistical Society*, 48(3):259–302, 1986.

[6] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *In Proc. 11th Annual Conf. Computational Learning Theory*, pages 92 – 100, 1998.

[7] J. W. Boardman. Geometric mixture analysis of imaging spectrometry data. In *Proc. Int. Geosci. Rem. Sens. Symp.*, pages 2369 – 2371, 1994.

[8] L. Breiman. Bagging Predictors. *Machine Learning*, 24(2):123–140, 1996.

[9] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[10] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.

[11] G. Camps-Valls and L. Bruzzone. Kernel-based methods for hyperspectral image classification. *IEEE Trans. Geosci. and Remote Sens.*, 43(6):1351–1362, Jun. 2005.

[12] G. Camps-Valls, L. Gomez-Chova, J. Calpe-Maravilla, J. D. Martin-Guerrero, E. Soria-Olivas, L. Alonso-Chorda, and J. Moreno. Robust support vector method for hyperspectral data classification and knowledge discovery. *IEEE Trans. Geosci. and Remote Sens.*, 42(7):1530–1542, Jul 2004.

[13] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-Frances, and J. Calpe-Maravilla. Composite kernels for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, 3(1):93 – 97, 2006.

[14] Y. Chen, M. M. Crawford, and J. Ghosh. Integrating support vector machines in a hierarchical output decomposition framework. In *2004 International Geosci. and Remote Sens. Symposium*, pages 949–953, Anchorage, Alaska, Sept. 20-24 2004.

[15] Y. Chen, M. M. Crawford, and J. Ghosh. Applying nonlinear manifold learning to hyperspectral data for land cover classification. In *2005 International Geosci. and Remote Sens. Symposium*, Seoul, South Korea, Jul. 24-29 2005.

[16] C. Cortes and V. Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.

[17] M. Crawford, J. Ham, and J. Ghosh. Robust classifiers for hyperspectral data analysis using limited training data. In *The 2003 Tyrrhenian International Workshop on Remote Sensing*, pages 15 – 18, Elba Island, Italy, September 2003.

[18] M. M. Crawford, J. Ham, Y. Chen, and J. Ghosh. Random Forests of Binary Hierarchical Classifiers for Analysis of Hyperspectral Data. In *Advances in Techniques for Analysis of Remotely Sensed Data, 2003 IEEE Workshop on*, pages 337– 345, 2003.

[19] M.M. Crawford, S. Kumar, M.R. Ricard, J.C. Gibeaut, and A. Neuenschwander. Fusion of airborne polarimetric and interferometric SAR for classification of coastal environments. *IEEE Transactions on Geoscience and Remote Sensing*, 37(3):1306–1315, 1999.

[20] J. A. Dare. Support Vector Machines in a Binary Hierarchical Classifier. Master's thesis, University of Texas at Austin, 2004.

[21] W. A. Davis and F. G. Peet. A method of smoothing digital thematic maps. *Remote Sensing of Environment*, 6:45–49, 1977.

[22] V. de Silva and J. B. Tenebaum. *Advances in Neural Information Processing System*, chapter Global versus local methods in nonlinear dimensionality reduction, pages 705–712. MIT Press, 2002.

[23] H. Derin and H. Elliott. Modeling and segmentation of noisy and textured images using gibbs random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(1):39–55, Jan 1987.

[24] Thomas G. Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.

[25] E. W. Dijkstra. Note on two problems in connection with graphs. *Numberische Mathematik*, 1:269–271, 1959.

[26] R. A. Fisher. The Statistical Utilization of Multiple Measurements. In *Annals of Eugenics*, volume 8, pages 378–386, 1938.

[27] G. M. Foody and A. Mathur. A relative evaluation of multiclass image classification by support vector machines. *IEEE Trans. Geosci. and Remote Sens.*, 42(6):1335–1343, Jun 2004.

[28] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. Technical report, Dept. of Statistics, Stanford University, 1998.

[29] J. Furnkranz. Round robin classification. *J. Machine Learning Research*, 2:721 – 747, 2002.

[30] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell*, 6(6):721–741, 1982.

[31] M. X. Goemans and D. P. Williamson. Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems Using Semidefinite Programming. *J. Assoc. Comput. Mach.*, 42:1115–1145, 1995.

[32] J. C. Gower and G. J. S. Ross. Minimum spanning trees and single linkage cluster analysis. *Applied Statistics*, 18:54–64, 1969.

[33] A. A. Green, M. Berman, P. Switzer, and M. D. Craig. A transformation for ordering multispectral data in terms of image quality with implications for noise removal. *IEEE Trans. on Geosci. and Remote Sens.*, 26(1):65–74, 1988.

[34] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh. Investigation of the random forest framework for classification of hyperspectral data. *IEEE Trans. Geosci. and Remote Sens.*, 43(3):492–501, Mar 2005.

[35] R. M. Haralick, K. Shanmugam, and Dinstein I. Textural features for image classification. *IEEE Trans. On SMC*, 3(6):610–621, 1973.

[36] A. Henneguelle, J. Ghosh, and M. M. Crawford. Polyline Feature Extraction for Land Cover Classification using Hyperspectral Data. In *1st Indian International Conference on Artifical Intelligence*, 2003.

[37] T. K. Ho. The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.

[38] C.-W. Hsu and C.-J. Lin. A comparison on methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13:415–425, 2002.

[39] Q. Jackson and D. Landgrebe. An adaptive classifier design for high-dimensional data analysis with a limited training data set. *IEEE Trans. Geosci. Rem. Sens.*, 39(12):2664 – 2679, 2001.

[40] Qiong Jackson and David Landgrebe. Adaptive bayesian contextural classification based on markov random fields. *IEEE Transactions on Geoscience and Remote Sensing*, 40(11):2454–2463, Nov 2002.

[41] X. Jia and J. A. Richards. Segmented principal components transformation for efficient hyperspectral remote-sensing image display and classification. *IEEE Trans. Geosci. Rem. Sens.*, 37(1):538–542, 1999.

[42] Luis O. Jimnez, Jorge L. Rivera-Medina, Eladio Rodrguez-Daz, Emmanuel Arzuaga-Cruz, and Mabel Ramrez-Vlez. Integration of spatial and spectral in-

formation by means of unsupervised extraction and classification for homogenous objects applied to multispectral and hyperspectral data. *IEEE Trans. Geosci. and Remote Sens.*, 43(4):844–851, Apr 2005.

[43] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, 1998.

[44] D. Korycinski. *Investigating the use of Tabu Search to find near-optimal solutions in multiclassifier systems*. PhD thesis, University of Texas at Austin, 2003.

[45] S. Kullback. *Information Theory and Statistics*. John Wiley and Sons., New York, 1959.

[46] S. Kumar. *Modular Learning through Output Space Decomposition*. PhD thesis, University of Texas at Austin, 2000.

[47] S. Kumar and J. Ghosh. Gamls: A generalized framework for associative modular learning systems. In *Application and Science of Computational Intelligence II, SPIE*, volume 3722, pages 24–35, Orlando, FL., April 1999.

[48] S. Kumar, J. Ghosh, and M. M. Crawford. Best-bases feature extraction algorithms for classification of hyperspectral data. *IEEE Trans. Geosci. Remote Sens.*, 39:1368–1379, 2001.

[49] S. Kumar, J. Ghosh, and M.M. Crawford. Hierarchical Fusion of Multiple Classifiers for Hyperspectral Data Analysis. *International J. Pattern Analysis and Applications*, 5(2):210–220, 2002.

[50] L. I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley, Hoboken, NJ, 2004.

[51] David Landgrebe. Hyperspectral image data analysis as a high dimensional signal processing problem. *Special Issue of the IEEE Signal Processing Magazine*, 19(1):17 – 28, 2002.

[52] C. Lee and D. A. Landgrebe. Decision boundary feature extraction for neural networks. *IEEE Trans. Neural Networks*, 8(1):75–83, Jan. 1997.

[53] S. Lee and M. M. Crawford. Unsupervised multistage image classification using hierarchical clustering with a bayesian similarity measure. *IEEE Trans. on Image Processing*, 14:312–320, 2005.

[54] F. Melgani and L. Bruzzone. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Trans. Geosci. and Remote Sens.*, 42(8):1778 – 1790, Aug. 2004.

[55] J. T. Morgan. *Adaptive Hierarchical Classification with Limited Training Data*. PhD thesis, University of Texas at Austin, 2001.

[56] J. T. Morgan, A. Henneguelle, M. M. Crawford, and J. Ghosh. Adaptive feature spaces for land cover classification with limited ground truth. *Intl. J. Pattern Recognit. Artif. Intell.*, 18(5):777–800, 2004.

[57] Y. Nesterov and A. Nemirovskii. *Interior Point Polynomial Methods in Convex Programming: Theory and Applications*. Society for Industrial and Applied Mathematics, Philadelphia, 1994.

[58] A. L. Neuenschwander, M. M. Crawford, and S. Ringrose. Results of the eo-1 experiment - use of earth observing-1 advanced land imager (ali) data to assess the vegetational response to flooding in the okavango delta, botswana. *Int. J. Remote Sens.*, 26(19):4293–4319, Oct 2005.

[59] J. S. Pearlman, P. S. Berry, C. C. Segal, J. Shapanski, D. Beiso, and S. L.

Carman. Hyperion: a space-based imaging spectrometer. *IEEE Trans. Geosci. Rem. Sens.*, 41(6):1160 – 1173, 2003.

[60] S. Poljak and Z. Tuza. *Maximum cuts and largest bipartite subgraphs*, pages 181–244. Mathematical Society, Providence, RI,, 1995.

[61] J. R. Quinlan. *C4.5 Programs for Machine Learning*. Morgan Kaufmann, 1988.

[62] S. Rajan. Empirical Evaluation of ECOC vs. BHC Approaches to Multi-Class Classification Problems. Master's thesis, University of Texas at Austin, 2004.

[63] S. Rajan and J. Ghosh. An Empirical Comparison of Hierarchical vs. Two-level approaches to Multiclass Problems. In F. Roli, J. Kittler, and T. Windeatt, editors, *Multiple Classifier Systems*, pages 283–292. LNCS Vol. 3077, Springer, 2004.

[64] S. Rajan, J. Ghosh, and M. M. Crawford. Exploiting class hierarchies for knowledge transfer in hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 44(11):3408 – 3417, 2006.

[65] S. Rajan, J.Ghosh, and M. M. Crawford. An active learning approach to knowledge transfer for hyperspectral data analysis. In *IEEE International Geoscience and Remote Sensing Symposium*, Denver, Colorado, August 2006.

[66] S. J. Raudys and A. K. Jain. Small sample size effects in statistical pattern recognition: recommendations for practitioners. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(3):252 – 264, 1991.

[67] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by local linear embedding. *Science*, 290(5500):2323–2326, 2000.

[68] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. In *Proc. 14th Inter-*

national Conference on Machine Learning, pages 322–330. Morgan Kaufmann, 1997.

[69] Bernhard Schlkopf, Alexander Smola, and Klaus-Robert Mller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, July 1998.

[70] G. A. F. Seber. *Multivariate Observation*. John Wiley & Sons, 1984.

[71] S. B. Serpico and L. Bruzzone. A new search algorithm for feature selection in hyperspectral remote sensing images. *IEEE Trans. Geosci. Rem. Sens.*, 39(7):1360 – 1367, 2001.

[72] M. Skurichina and R. Duin. Bagging, Boosting and the Random Subspace Method for Linear Classifiers. *Pattern Analysis and Applications*, 5:121–135, 2002.

[73] S. Tadjudin and D. A. Landgrebe. Covariance estimation with limited training samples. *IEEE Trans. Geosci. Remote Sens.*, 37(4):2113–2118, Jul. 1999.

[74] Y. Takane, F. W. Young, and J. De Leeuw. Non-metric Individual Differences Multidimensional Scaling: Alternating Least Squares with Optimal Scaling Features. *Psycometrika*, 42:7–67, 1977.

[75] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[76] M. J. Todd. Semidefinite Optimization. Technical report, Cornell University, Ithaca, NY 14853, 2001.

[77] K. Tumer and J. Ghosh. Error Correlation and Error Reduction in Ensemble Classifiers. *Connection Science*, 8(3-4):385–403, 1996.

[78] L. A. Wolsey. *Integer Programming.* John Wiley & Sons, 1998.

Vita

Yang-Chi Chen was born in Taichung, Taiwan on November 2, 1973, the son of Ya-Yi Chen and Chi-Ru Yang. He received the B.S. in Naval Architecture and Ocean Engineering from National Taiwan University, Taipei, Taiwan in 1996. In September 1999 he entered the Graduate School of The University of Texas at Austin. He received his M.S. in Operations Research and Industrial Engineering from The University of Texas at Austin in 2001 and continued for his Ph.D. degree in the same program. He worked at The University of Texas Center for Space Research as a Graduate Research Assistant from June 2002 to December 2005. He is currently a Senior Project/Program Analyst at Advanced Micro Devices (AMD). He has published six conference papers and one journal paper during his study. He also served as the President of Taiwanese Student Association at The University of Texas at Austin from July 2001 to June 2002.

Permanent Address: 3373 Lake Austin Blvd.
Apt. A
Austin TX, 78703

This dissertation was typeset with LaTeX 2$_\varepsilon$[1] by the author.

[1] LaTeX 2$_\varepsilon$ is an extension of LaTeX. LaTeX is a collection of macros for TeX. TeX is a trademark of the American Mathematical Society. The macros used in formatting this dissertation were written by Dinesh Das, Department of Computer Sciences, The University of Texas at Austin, and extended by Bert Kay, James A. Bednar, and Ayman El-Khashab.